Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Visualização Exploratória e Predição de Desempenho de Dados Educacionais da Universidade de Brasília

Pedro Borges Pio, Igor Chaves Sodré

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Vinicius Ruela Pereira Borges

Brasília
2019

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Visualização Exploratória e Predição de Desempenho de Dados Educacionais da Universidade de Brasília

Pedro Borges Pio, Igor Chaves Sodré

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Vinicius Ruela Pereira Borges (Orientador)
CIC/UnB

Prof. Dr. Jorge Henrique Cabral Fernandes
CIC/UnB

Prof. Dr. Jan Mendonça Correa
CIC/UnB

Prof. Dr. Edison Ishikawa
Coordenador do Bacharelado em Ciência da Computação

Brasília, 29 de Agosto de 2019

Dedicatória

Aos nossos familiares e amigos.

Agradecimentos

Gostaríamos de agradecer ao nosso orientador Professor Doutor Vinicius Ruela Pereira Borges que nos auxiliou em durante todo o decorrer do trabalho, ajudando a escolher o tema, tirando duvidas e fazendo sugestões.

Pedro: Agradeço à minha família que sempre me apoiou durante toda a minha trajetória, aos meus amigos que melhoraram a minha experiência durante o período universitário e à minha dupla, que, assim como o nosso orientador, tornou possível a realização deste trabalho.

Igor: Gostaria de agradecer aos meus amigos e em especial ao ***master group*** por me manterem motivados durante a graduação.

Resumo

Nas últimas décadas, a quantidade dados educacionais coletados por universidades tem aumentado. Com o auxílio das informações passíveis de serem extraídas desses dados, universidades podem melhorar a qualidade de ensino e o desempenho dos estudantes. Técnicas de visualização de informação podem ser úteis para a análise desses dados facilitando a descoberta de padrões e auxiliando universidades a tomar decisões.

Em 2004, a Universidade de Brasília (UnB) começou a implementar um sistema de cotas raciais no ingresso dos estudantes, separando parte das vagas para estudantes negros. A quantidade de dados gerados, referentes a esses estudantes nos últimos 15 anos, torna possível realizar uma comparação entre os estudantes cotistas de forma quantitativa e verificar diferenças em seus rendimentos acadêmicos.

Neste trabalho, por meio da visualização exploratória foi realizada uma comparação entre os alunos cotistas e não cotistas da UnB. Técnicas de redução de dimensionalidade como PCA e t-SNE foram utilizadas para facilitar a visualização e separação dos estudantes em grupos. Uma ferramenta de visualização interativa também foi construída para prover informações de acordo com a necessidade do usuário. Além disso, foi desenvolvido um modelo de predição, utilizando os algoritmos *K-NN* e *Gradient Boosting*, para prever os alunos que estão com maior risco de abandonar o curso, provendo assim informações para a universidade, se possível.

Palavras-chave: Visualização Exploratória, Dados Educacionais, Ações Afirmativas, Aprendizado de Máquina

Abstract

In recent decades the amount of educational data collected by universities has increased. With the help of information that can be extracted from this data, universities can improve teaching quality and student performance. Information visualization techniques can be useful for analyzing this data by facilitating pattern discovery and helping universities make decisions.

In 2004, the University of Brasilia (UnB) began implementing a quota system for student admission, separating part of the vacancies for black students. The amount of data generated for these students over the last 15 years makes it possible to perform a quantitative comparison between quota students and verify differences in their academic achievement.

In this study, through visual data mining, a comparison between the quota and non-quota students was performed at UnB. Dimensionality reduction techniques such as PCA and t-SNE were used to facilitate visualization and separation of students into groups. An interactive visualization tool has also been built to provide information according to the user's need. In addition, a prediction model has been developed using the K-NN and Gradient Boosting to predict students who are most likely to drop out, thus providing information to the university and, if possible, to avoid dropout.

Keywords: Visual Data Mining, Educational Data, Afirmative Actions, Machine Learning

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

Até o final do século XX, parte dos dados registrados nas universidades ainda eram armazenados em papéis. A transição recente deste meio físico de armazenamento para bancos de dados informatizados influenciou no surgimento tardio de técnicas de mineração de dados na área educacional [1]. De forma análoga, com a maior quantidade de informações armazenadas, nos últimos anos foi observado um crescimento do volume de dados digitais, gerando problemas com relação à sua complexidade. Acredita-se que um dos principais fatores desse crescimento foi a utilização do computador para auxiliar as tarefas diárias, tanto *online* quanto *offline* [2].

Entende-se por dados educacionais qualquer informação coletada sobre educadores, escolas, organizações de ensino e estudantes incluindo dados pessoais (e.g. raça, idade, endereço), informação de matrícula (e.g. nome da escola, período corrente, número de faltas), informação acadêmica (e.g. disciplinas cursadas, notas em testes) além de outros tipos variados de dados (e.g. notas disciplinares, problemas de saúde) [3, 4]. Adicionalmente, sistemas de ensino podem gerar dados educacionais intrínsecos ao seu próprio uso (e.g. tempo de resposta de uma questão, geolocalização do computador utilizado), o que aumenta bastante a variedade de dados que provêm informação útil [5]. Esta grande quantidade de atributos e a falta de padronização acarreta numa dificuldade da análise significava sem o auxílio de técnicas de manipulação e visualização de dados, uma vez que a capacidade de analisar valores tabulares sem ferramentas visuais é severamente restrita. Estas informações extraídas dos dados educacionais podem ser de grande auxílio a universidades e outros órgãos de ensino de diversas maneiras como:

- Apoiam estudantes a otimizarem o estilo de aprendizagem e o material de ensino como, por exemplo, alterar a razão entre tempo recebendo aula e tempo fazendo exercícios, além de detectar quais exercícios fornecem uma melhor compreensão dos conteúdos por parte do aluno [2].

- Apoiam educadores em tarefas de análise do comportamento do estudante e ajudam a prever seu aprendizado para aumentar a efetividade do ensino [2].
- Apoiam pesquisadores a avaliar o material de ensino e melhorar sistemas educacionais [2].
- Juntamente com técnicas de aprendizado de máquina, podem prever performance e evasão dos alunos de maneira que ações podem ser tomadas de antemão para evitá-la [2].

Sempre que uma universidade se esforça para que mulheres, homens, negros, brancos e portadores de necessidades especiais tenham as mesmas oportunidades de receberem educação, esta organização possui uma política de ação afirmativa [6]. A implementação de políticas de ações afirmativas se deu recentemente na Universidade de Brasília (UnB), em 2004, sendo a primeira universidade brasileira a reservar 20% das suas vagas para alunos cotistas [6]. Os alunos beneficiários de ações afirmativas são uma parcela dos estudantes que geram dados educacionais importantes de serem investigados para que possamos entender se existem diferenças de performance, evasão, entre outros em relação aos outros alunos e, caso exista, propor ações que podem ser tomadas para que este quadro seja melhorado.

Um ponto importante a ser investigado através de dados educacionais é a evasão, que é caracterizada pela saída do estudante de um dos cursos ou da instituição de ensino de maneira temporária ou definitiva e é considerada um dos problemas mais graves do ensino superior [7]. Um estudo da UnB, mostra que entre os anos de 2002 e 2008, 23.9% dos estudantes da universidade não se formaram e somente 29.5% se formaram no prazo [8]. Nota-se que nem todas as instituições de ensino brasileiras utilizam medidas de para evitar à evasão [7], o que motiva este estudo a focar em partes neste problema.

A maneira como informação é apresentada influencia na sua compreensão. Tal fato vale para a apresentação de dados educacionais. Aproveitando que seres humanos têm uma facilidade maior de extrair e assimilar informação de estímulos visuais dinâmicos (e.g. gráficos, imagens, vídeos) em comparação com estímulos visuais textuais (e.g. artigos, livros, tabelas) [9], a prática de visualização exploratória para descoberta de conhecimento implícito aos dados tem se mostrado um recurso vantajoso [10]. Alguns exemplos da aplicação deste recurso podem ser vistos em vários campos da educação tais quais: educação tradicional [11, 12, 13, 14]; *e-learning* [15, 16, 17, 18]; e sistemas inteligentes [19, 20, 21].

Com isso, nota-se que técnicas de visualização exploratória podem ser muito úteis para auxiliar universidades, visto que algumas coordenações de cursos carecem de ferramentas de descoberta de conhecimento implícito para os dados de seus estudantes. Assim, este estudo visa suprir parte desta demanda por meio de uma visualização exploratória dos dados educacionais dos alunos da UnB.

1.2 Objetivos

Com maior foco no auxílio e melhoria no desempenho dos estudantes, utilizando técnicas de visualização de dados, aprendizado de máquina e visualização exploratória, este trabalho, motivado pelos problemas de evasão nas instituições de ensino brasileiras e da carência de ferramentas de visualização exploratória focadas em dados educacionais, possui três objetivos principais:

- Comparar o desempenho e evasão de estudantes cotistas e não cotistas com o intuito de verificar possíveis padrões entre esses alunos por meio de técnicas de visualização de dados;
- Elaborar um modelo preditivo para descobrir se um aluno concluirá ou não o curso, possibilitando a instituição tomar providências para prevenir a evasão;
- Aplicação de visualizações interativas de dados para a uma análise definida pelo usuário e a descoberta de padrões nos dados dos estudantes.

Acredita-se que, cumprindo os objetivos propostos, as universidades terão mais informações para aprimorar a qualidade de ensino dos estudantes, tanto cotistas quanto não cotistas, além de possibilitar a prevenção da evasão de estudantes. Adicionalmente, espera-se que a utilização da ferramenta construída para a visualização interativa possibilite a descoberta de conhecimento implícitos aos dados dos alunos do ensino superior.

1.3 Estrutura do Documento

O restante deste trabalho é dividido em quatro capítulos:

- Fundamentação Teórica: São introduzidos conceitos básicos de estudos de dados, fundamentos de estatística, técnicas de visualização da informação e aprendizado de máquina. Todos necessários para uma maior compreensão do trabalho realizado.
- Visualização exploratória de alunos cotistas e previsão de desempenho: É apresentado um estudo sobre os alunos cotistas dos cursos do departamento de Ciência da Computação da UnB juntamente com um modelo de previsão indicando se os estudantes do departamento concluirão, ou não, o curso.
- Visualização exploratória de dados educacionais: É desenvolvida uma forma de visualização de informação interativa aplicada nos dados dos estudantes com um intuito de facilitar a extração e o entendimento dos dados estudados.
- Conclusão: São exibidas as conclusões finais sobre o trabalho juntamente com possíveis trabalhos futuros a serem estudados.

Capítulo 2

Fundamentação Teórica

Com o objetivo de facilitar o entendimento do trabalho, este capítulo, introduz a fundamentação teórica sobre o que será abordado nos capítulos seguintes. Este capítulo foi dividido em 5 partes: a primeira define conceitos básicos sobre dados, explicando o modelo tabular, instâncias, atributos e dimensionalidade; a segunda introduz conceitos estatísticos utilizados, como média, variância e matriz de correlação; a terceira apresenta formas de visualização de dados e de redução de dimensional; a quarta aborda aprendizado de máquina, mostrando modelos de classificação e de agrupamento de dados; a quinta cita considerações finais sobre o capítulo.

2.1 Estudo de Dados

2.1.1 Definições Utilizadas

Medida

Quantidade fixada por um padrão para determinar as dimensões ou o valor de uma grandeza da mesma espécie.

Conjunto de dados *Iris*

Conjunto de dados multidimensional gerado pelo biólogo Ronald Fisher [22] foi utilizado para exemplificar algumas técnicas de visualização. Este conjunto possui 3 tipos de flores (***Iris setosa, Iris virginica e Iris versicolor***) em que foram medidos quatro atributos para cada uma das amostras totalizando 150 flores (50 para cada tipo):

1. *Sepal length*: Comprimento da sépala
2. *Sepal width*: Largura da sépala
3. *Petal length*: Comprimento da pétala

4. *Petal width*: Largura da pétala

2.1.2 Modelo Tabular, Instância e Atributo

Para propósitos de formalização, no decorrer deste estudo, um conjunto de dados multidimensional será definido por $\mathbf{X} : \{x_1, x_2, x_3, ..., x_n\}$ em que x_i é uma instância, caracterizada por *d* atributos $x_i = \{x_{i1}, x_{i2}, x_{i3}, ..., x_{id}\}$.

O modelo tabular é uma estrutura de dados formatada e organizada em linhas e colunas que expressam relações e metadados, i.e., informações descritivas sobre tais relações como os rótulos das linhas e colunas [23]. Neste modelo, para representar um conjunto de dados *d*-dimensional, cada instância se associa a uma linha da tabela e cada coluna se associa aos valores de determinado atributo. Além disso, todos os valores presentes em uma coluna possuem o mesmo tipo, o que torna possível a comparação entre as instâncias. A Tabela 2.1 exemplifica esse modelo, em que cada linha representa uma instância do conjunto de dados, que tem o identificador do aluno (ID Aluno), Período de Ingresso, Média do Período e Nome da Disciplina como os atributo dos dados.

ID Aluno	Período de Ingresso	Média do Período	Nome da Disciplina
99975	20102	4	Teoria dos Números 1
993545	20102	3.5	Física 1
99985	20101	4.5	Álgebra 1
99973	20112	3	Cálculo 3
100654	20121	3	Noções de Direito

Tabela 2.1: Exemplo do modelo tabular contendo quatro atributos e cinco instâncias.

A grande quantidade de atributos presentes nos dados educacionais utilizados nesta pesquisa, torna inviável a sua análise utilizando somente a visualização de dados em tabelas. Mesmo assim, esta prática é imprescindível para que se possa ter um entendimento preliminar do tipo dos dados, além das ações a serem propostas na fase de pré-processamento.

2.1.3 Dissimilaridade

A comparação entre instâncias de um conjunto de dados é uma operação útil para sua análise. Por isso, formas de calcular o grau de semelhança são fundamentais. A dissimilaridade pode ser descrita como um valor numérico variando no intervalo $[0, \infty)$ para mostrar o quão diferentes entre si são duas instâncias de um conjunto de dados, em que valores próximos de 0 indicam que são parecidas [24]. Medidas de dissimilaridade são essenciais para problemas de reconhecimento de padrões como classificação e agrupamento de dados [24]. Uma das maneiras de

calculá-la é utilizando funções de distância. Seguem nas próximas sub-seções as definições de algumas destas funções mais conhecidas.

2.1.4 Distância Euclidiana

Seja um conjunto ***X***, a distância Euclidiana para cada par (x_i, x_j) é descrita pela Eq. (2.1) [24]:

$$d_E(i,j) = (\sum_{k=1}^{d}(x_{ik} - x_{jk})^2)^{\frac{1}{2}}. \tag{2.1}$$

2.1.5 Distância de Minkowski

A distância de Minkowski é uma generalização da distância Euclidiana. Considerando um conjunto ***X***, a distância de Minkowski entre a i-ésima e a j-ésima instância é descrita pela Eq. (2.2) [24]:

$$d_M(i,j) = (\sum_{k=1}^{d}|x_{ik} - x_{jk}|^{\lambda})^{\frac{1}{\lambda}}, \tag{2.2}$$

em que $\lambda \geq 1$. Sendo também chamada de L_λ.

2.2 Fundamentos de Estatística

A escolha correta do ferramental estatístico que será utilizado é um passo de extrema importância. Essas ferramentas são utilizadas como base para as técnicas de redução de dimensionalidade e de aprendizado de máquina, além disso, são úteis para a extração de informações e análise de dados. Segue abaixo as definições utilizadas neste estudo.

2.2.1 Média Aritmética

As medidas de tendência central guarnecem um valor numérico representativo de uma distribuição de valores. A média aritmética de um conjunto ***X*** unidimensional de tamanho *n* é definida pela Eq. (2.3) [25]:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.3}$$

2.2.2 Variância

O valor médio não provêm toda informação essencial para um conjunto de medidas. A variância é utilizada para medir o grau em que os dados numéricos tendem a se dispersar em torno da média do conjunto, proporcionando assim, um nível de confiabilidade. A variância de um conjunto de medidas **X** de tamanho *n* é definida pela Eq. (2.4) [25]:

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2, \tag{2.4}$$

em que $\bar{X}$ é média do conjunto dado pela Eq. (2.3).

2.2.3 Covariância

A covariância, descrita pela Eq. (2.5), é uma forma de medir grau de associação entre duas variáveis de um conjunto de dados.

$$Cov(X, Y) = E(XY) - E(X)E(Y), \tag{2.5}$$

em que $E(X)$ é o valor esperado de uma variável.

Quando a covariância $Cov(X, Y)$ é positiva e, x_i for maior que a média do conjunto de variáveis da qual pertence, y_i tende a ser maior que sua média. Analogamente, quando a covariância é negativa e, x_i for maior que a média do conjunto de variáveis da qual pertence, y_i tende a ser menor que sua média [26].

2.2.4 Correlação

Dois atributos são considerados dependentes positivamente quando o aumento do valor de um corresponde ao aumento do valor do outro. Analogamente, são considerados dependentes negativamente quando o aumento do valor de um corresponde ao decremento no valor do outro. Uma das maneiras de medir tal dependência é a correlação de Pearson [27], cujo valor varia no intervalo $[-1, 1]$. Considerando dois conjuntos *X* e *Y*, o coeficiente de correlação é definido pela Eq. (2.6):

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{Y})^2}}, \tag{2.6}$$

em que $\bar{X}$ e $\bar{Y}$ é a média aritmética definida na Eq. (2.3).

Pode se definir uma matriz de correlações como uma matriz em que cada posição representa o coeficiente de correlação entre duas variáveis definidas na Eq. (2.6). A matriz de correlação, exemplificada na Tabela 2.2, pode ser usada para resumir um conjunto de dados ou como entrada para uma análise mais avançada (e.g. construção de *heat maps*) [28].

	B1	B2	B3	B4	B5
B1	1				
B2	0.53	1			
B3	0.73	0.44	1		
B4	0.87	0.96	0.41	1	
B5	0.43	0.71	0.72	0.56	1

Tabela 2.2: Exemplo de matriz de correlação.

2.3 Visualização da Informação

Na literatura, uma das definições mais utilizadas para visualização da informação é "A utilização de um auxílio computacional interativo em uma representação visual dos dados para ampliar a cognição", em que cognição é o poder de percepção humana [29]. Aplicações de visualização são comumente utilizadas para criar tabelas, imagens e outras formas intuitivas de representação de dados [30]. Considerando as diferentes aplicações de visualização de dados, é natural que se crie diferentes técnicas para a extração de informações.

2.3.1 Visualização Exploratória

Quando a visualização de dados é feita em forma textual, a amostra passível de análise humana é ínfima em relação ao grande volume de dados disponível. O processo de visualização exploratória é especialmente útil quando se sabe muito pouco sobre a estrutura dos dados e os objetivos de pesquisa ainda estão vagamente definidos. Outros usos bastante proeminentes incluem a formação e verificação de hipóteses [10].

Existem três etapas distintas que são usualmente utilizadas no processo de visualização exploratória: visão geral dos dados, filtragem e exibição de detalhes. Essas etapas são apoiadas por um interlocutor humano, que introduz flexibilidade, criatividade e interatividade ao processo. Na visualização geral dos dados, o foco se dá na identificação de padrões e grupos presentes que sejam perceptíveis em uma análise superficial em todo o conjunto. Na filtragem, o foco se move para um subconjunto específico dos dados. Por fim, na exibição de detalhes, são mostradas informações sobre o subconjunto selecionado na etapa de filtragem. Técnicas de visualização de dados podem ser utilizadas para todas as etapas descritas previamente. Vale lembrar que a visualização exploratória lida facilmente com dados heterogêneos e ruidosos, além de não requerer entendimento prévio de construções matemáticas e estatísticas complexas [10].

2.3.2 Técnicas de Visualização Clássicas

Uma decisão importante no processo de visualização exploratória é a escolha das técnicas de visualização. Dependendo do objetivo da exploração ou do tipo de dado, certas técnicas são mais adequadas que outras, assim, facilitando a análise do usuário. Dessa forma, seguem abaixo algumas técnicas comuns de visualização.

Gráfico de Dispersão

Um gráfico de dispersão é a representação gráfica dos dados nas coordenadas cartesianas. Nele é mostrada a relação entre duas variáveis, a primeira (variável independente) representa a distância horizontal e a segunda (variável dependente) a distância vertical a partir do eixo de coordenadas [29].

Um exemplo de um Gráfico de dispersão é mostrado na Figura 2.1, em que é apresentada a visualização das variáveis *Sepal length* como variável independente e *Sepal width* como variável dependente, ambas pertencentes ao conjunto de dados *Iris*. Cada ponto neste gráfico representa uma flor com o comprimento e largura das sépalas indicadas pela sua coordenada cartesiana do gráfico.

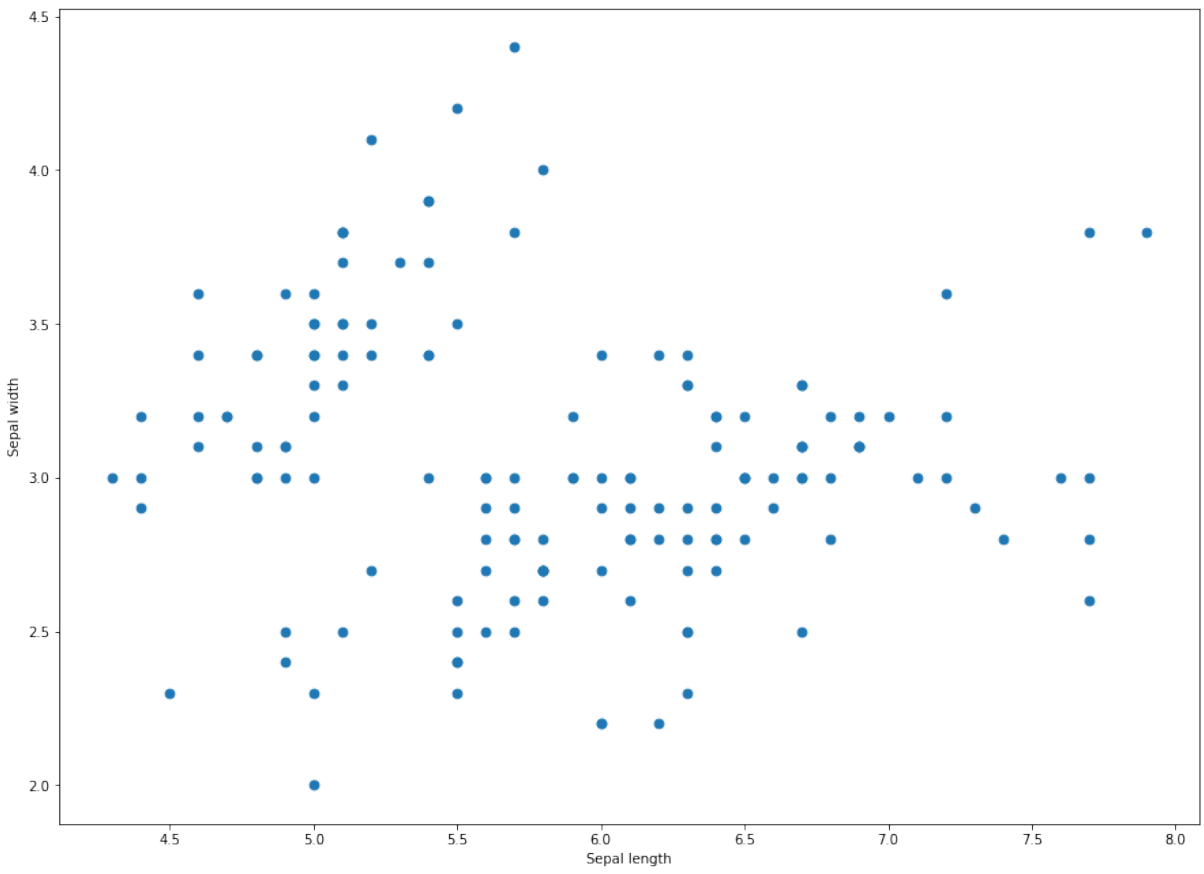


Figura 2.1: Exemplo de aplicação de um gráfico de dispersão.

Gráfico de Barras

Um dos métodos mais utilizados de visualização, o gráfico de barras, é comumente utilizado para representar dados discretos [29]. Neste tipo de gráfico, a magnitude de uma determinada categoria é agrupada e representada como um conjunto único por meio de uma barra. A Figura 2.2 exemplifica um gráfico de barras da quantidade de vinhos que receberam determinada nota [1].

[1] O conjunto de dados *Wine Review* utilizado para criação da Figura 2.2 foi obtido de `https://www.kaggle.com/zynicide/wine-reviews`

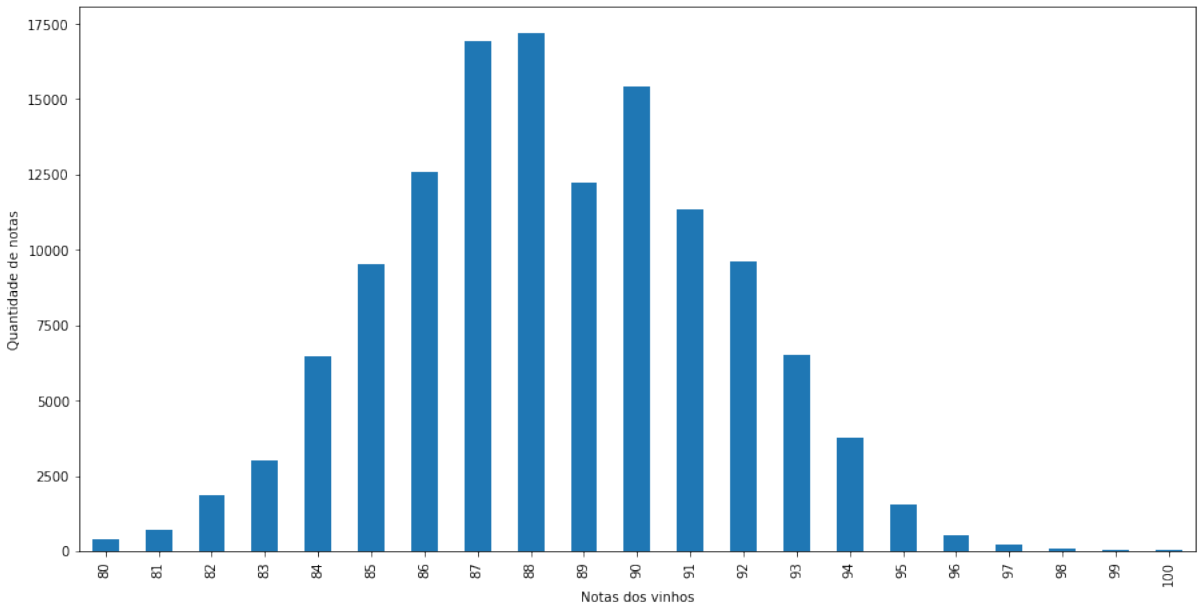


Figura 2.2: Exemplo de aplicação de um gráfico de barras.

Heat Map

Um *heat map* é uma representação gráfica de um dado em que os valores individuais contidos em uma matriz são representados com cores [31]. No *heat map*, uma instância de dados é apresentada quantitativamente em dois eixos, o *eixo-x* normalmente representa amostras individuais enquanto que o *eixo-y* consiste os parâmetros medidos. O espaço entre os eixos é composto de caixas coloridas de forma a refletir a quantidade da variável. Assim, *heat maps* são formas flexíveis de se visualizar grupos e explorar padrões nos dados [32].

A Figura 2.3 exemplifica um gráfico *heat map* de uma matriz de correlação das variáveis *sepal_length, sepal_width, petal_length, petal_width* do conjunto dados *Iris*.

Heat Map da Matriz de Correlação do Conjunto *Iris*

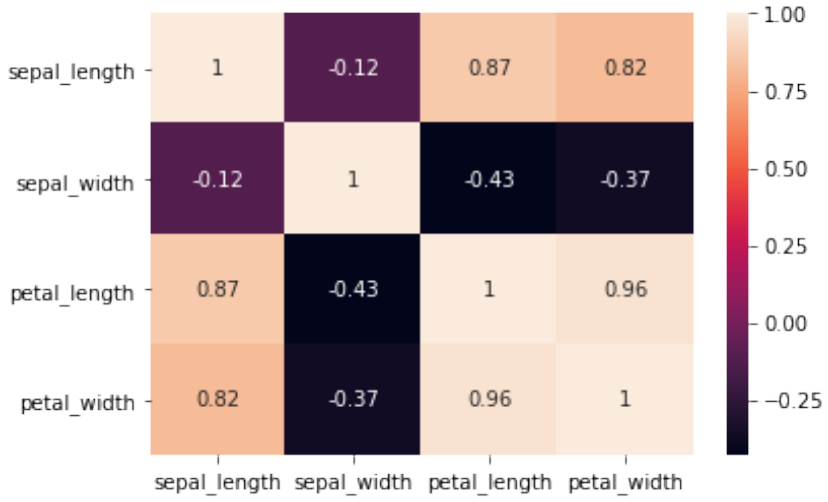


Figura 2.3: Exemplo de aplicação de um *heat map*.

Coordenadas Paralelas

O método de coordenadas paralelas representa uma instância d-dimensional como valores em d retas perpendiculares ao *eixo-x* igualmente espaçadas. Cada instância de dados d-dimensional é representado por uma polilinha entre as retas conectando-as com seu respectivo valor. Este método proporciona uma forma de visualizar dados de alta dimensionalidade em uma representação 2D, auxiliando a descoberta de padrões e correlações [33].

A Figura 3.8 exemplifica um gráfico de coordenadas paralelas, em que são representados quatro atributos do conjunto de dados *íris (sepal_length, sepal_width, Petal_lenght, Petal_width)* no *eixo-x*, e, cada polilinha representa uma instância de dados categorizadas em três classes diferentes *(Setosa, Versicolor e Virgínica)*. Nota-se, pelo gráfico, que somente as flores do tipo setosa possuem comprimento da pétala menor que 2 e largura da pétala menor que 1, além disso as flores da espécie virgínicas possuem sépalas e pétalas com maior comprimento.

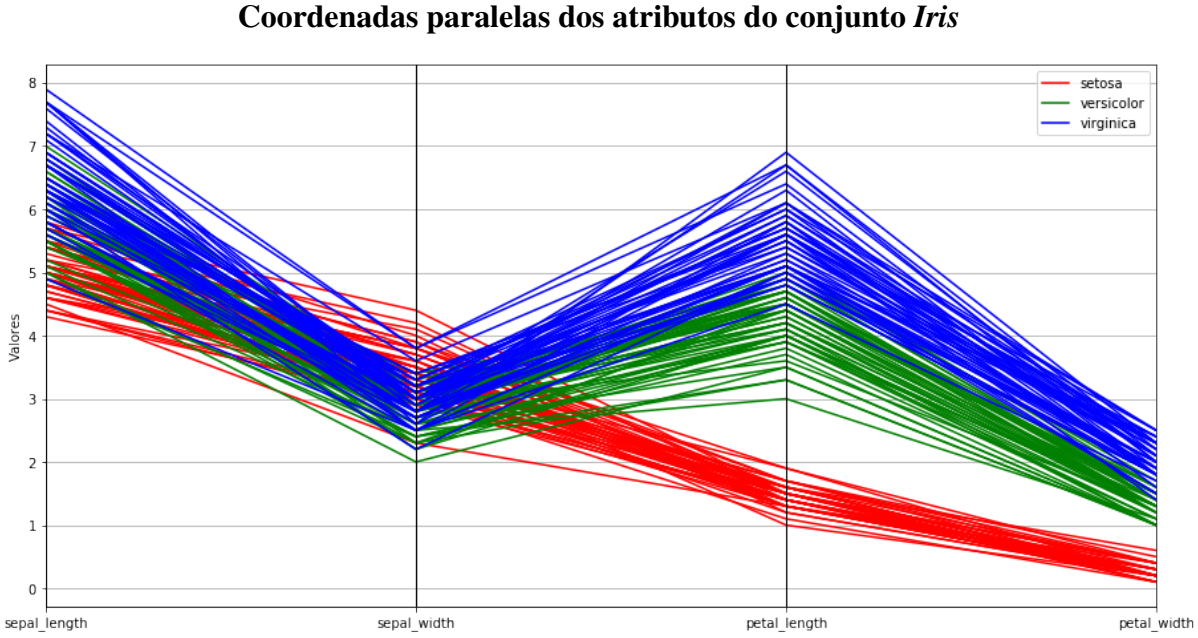


Figura 2.4: Exemplo de aplicação de coordenadas paralelas.

2.3.3 Técnicas de Visualização Baseadas em Redução de Dimensionalidade

Mesmo facilitando a visualização de dados e à extração de informações, as técnicas clássicas possuem limitações que normalmente estão ligadas a dimensionalidade dos dados. No caso dos gráficos de dispersão e de barras, o número de dimensões passíveis de serem representados ao mesmo tempo é bem limitado. Na visualização baseada em *Heat maps*, a medida que a dimensionalidade dos dados aumenta, a representação gráfica fica mais complexa e difícil de se interpretar devido ao grande número de linhas e colunas. Fato semelhante também ocorre nas visualizações baseadas em coordenadas paralelas, tendo sua interpretação afetada pela quantidade de instâncias do conjunto, gerando sobreposição das polilinhas e pela quantidade de atributos, gerando o aumento do tamanho do gráfico. Nota-se assim que a dimensionalidade dos dados é um fator importante para sua representação.

A maldição da dimensionalidade como descrita por Richard E. Bellman [34] se refere a vários fenômenos que surgem ao se analisar e organizar dados em espaços de alta dimensão que não ocorrem em baixas dimensões como o espaço físico tridimensional. O tema principal deste problema é que quando a dimensionalidade aumenta, o volume do espaço cresce rapidamente, a ponto de que os dados disponíveis se tornam espaçados. Esta dispersão é problemática para todo método que requer significância estatística. Para se obter resultados confiáveis, a quantidade de dados necessários para apoiar o resultado cresce exponencialmente com a dimensionalidade.

Com o intuito de evitar tais problemas, técnicas de redução de dimensionalidade podem ser úteis para visualização de informação. A redução de dimensionalidade pode ser feita tanto por seleção quanto por transformação de características. Na seleção de características, seleciona-se as variáveis a serem utilizadas por meio de algum critério definido, mantendo assim, os valores originais das variáveis. Na transformação de características, por meio de algum algoritmo, geram-se outros atributos para representar o conjunto de variáveis. Essas técnicas buscam encontrar uma representação de menor dimensão de dados multidimensionais em um espaço *p*-dimensional com $p = \{1, 2, 3\}$ de maneira a preservar uma parcela da informação de distância entre as instâncias [35].

O resultado de uma técnica de visualização baseada na redução de dimensionalidade é um conjunto de pontos no espaço visual podendo ser uma plano, reta ou volume. Preferencialmente, se pontos forem posicionados próximos neste espaço reduzido, significa que os itens que estes pontos representam são semelhantes de acordo com a dissimilaridade escolhida. Analogamente, se forem projetados distantes, indica que os itens representados são pouco relacionados [35]. Técnicas que mapeiam espaços multidimensionais em espaços visuais são ditas técnicas de posicionamento de pontos [35].

PCA

Principal Components Analysis (PCA) é uma técnica de redução de dimensionalidade linear, o que significa que ele diminui o número de dimensões dos dados incorporando-os em um subespaço linear de menor dimensão [36].

O PCA busca encontrar uma transformação ortogonal linear dos dados tal que sua máxima variação seja explicada pelas primeiras coordenadas no espaço transformado. O PCA define uma matriz de transformação quadrática ***W*** que, quando multiplicada por um elemento em ***X***, expressa este ponto em uma nova base ortogonal, em que os eixos são ordenados de maneira decrescente pela variância respectiva ao conjunto original. Utilizando os autovetores associados aos maiores autovalores de ***W***, cria-se uma transformação de redução de dimensionalidade ***W'***, que, quando multiplicada por um dos pontos x_i tem-se a relação:

$$\forall x_i \in X : y_i = x_i W' \tag{2.7}$$

em que y_i é a representação de baixa dimensão de x_i. A dimensionalidade do conjunto resultante ***Y*** corresponde ao número de colunas escolhido para a matriz ***W'*** [37].

As componentes principais do PCA são obtidas por decomposição espectral em valores singulares (SVD) da matriz de covariância do conjunto de dados [38]. Um exemplo de redução de dimensionalidade baseado no PCA é mostrado na Figura 2.5, em que o conjunto de dados

Iris de 4 dimensões é representado em um *layout* bidimensional. Pode-se dizer que as classes agora são linearmente separáveis, o que não era visível nas representações anteriores.

PCA Aplicado no Conjunto *Iris*

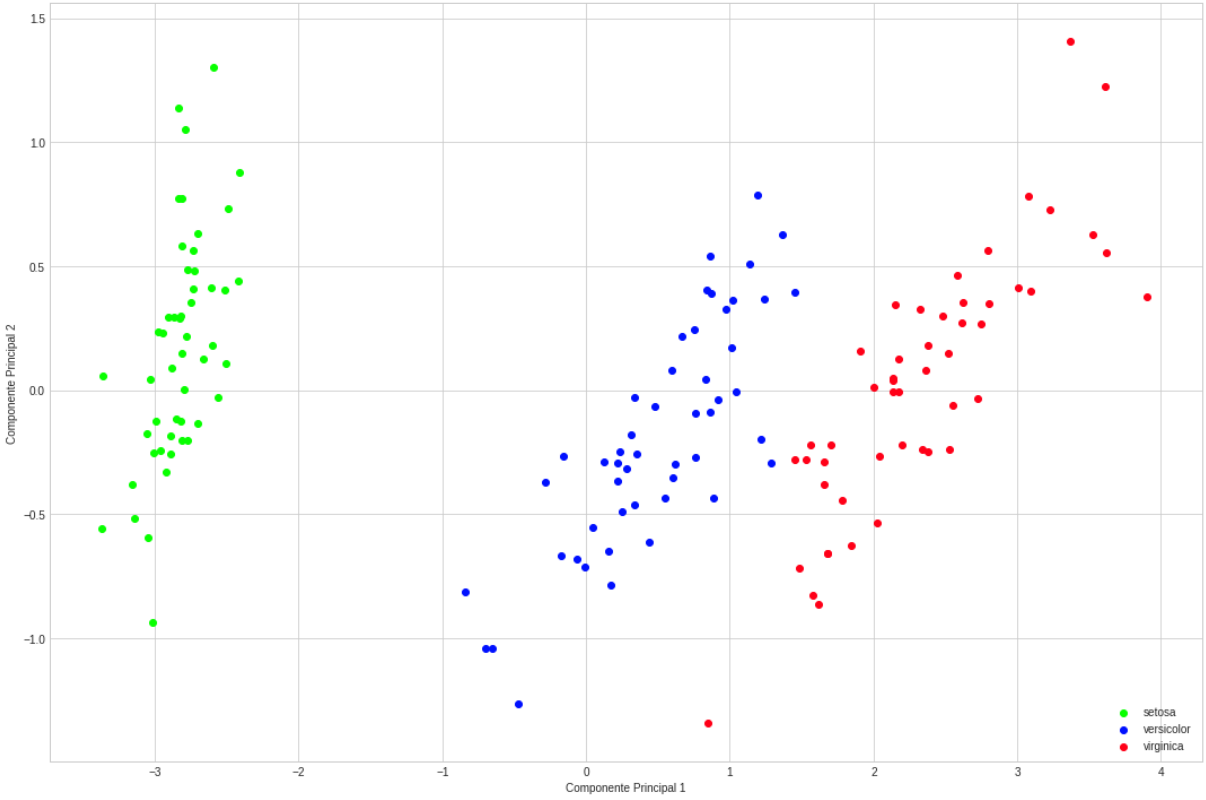


Figura 2.5: Exemplo de aplicação do PCA.

T-SNE

t-Distributed stochastic neighbor embedding (t-SNE) é uma técnica de redução de dimensionalidade baseada em posicionamento de pontos que minimiza a diferença entre duas distribuições: a que mede a semelhança entre os pares de objetos de entrada e a que mede a similaridade dos pares dos mesmos objetos projetados em um subespaço de menor dimensionalidade [39]. Dado um subconjunto multidimensional de objetos $\boldsymbol{X}$ e uma função $d(x_i, x_j)$ que calcula a distância entre um par de objetos, busca-se um subespaço S-dimensional que incorpore os pontos do conjunto $\boldsymbol{X}$ mapeados para novos pontos de um conjunto $\boldsymbol{Y}$ com $y_i \in \mathbb{R}^S$. Para este fim, t-SNE define probabilidades conjuntas p_{ij} que medem a similaridade de pares de objetos x_i e x_j, simetrizando duas probabilidades condicionais:

$$p_{j|i} = \frac{exp(\frac{-d(x_i,x_j)^2}{2\sigma_i^2})}{\sum_{k\neq i} exp(\frac{-d(x_i,x_k)^2}{2\sigma_i^2})}, \quad p_{i|i} = 0 \tag{2.8}$$

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n} \tag{2.9}$$

Na Eq. (2.8), a largura de banda do *Kernel* Gaussiano [40], σ_i, é definida de maneira que a perplexidade (medida efetiva do número de vizinhos de x_i) da distribuição condicional P_i se iguala a perplexidade predefinida u. Isso faz com que o valor ótimo para σ_i varie para cada objeto do conjunto, ou seja, para regiões do espaço com maior densidade de pontos, σ_i tende a possuir um menor valor em relação a regiões do espaço com menor densidades de pontos.

No conjunto S-dimensional $\boldsymbol{Y}$, a similaridade entre dois pontos y_i e y_j são calculados usando o *heavy-tailed kernel* normalizado [41]. Especificamente, a similaridade incorporada q_{ij} entre os pontos y_i e y_j é calculada como a *Student-t kernel* normalizada [42] com um único grau de liberdade conforme a Eq. (2.10).

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|y_{ik} - y_{jl}\|)^{-1}}, \quad q_{ii} = 0. \tag{2.10}$$

Este procedimento permite objetos de entrada dissimilares x_i e x_j serem modelados pelas suas projeção de baixa dimensão y_i e y_j mantendo essa dissimilaridade [39].

As localizações dos pontos y_i mapeados no espaço de baixa dimensão são determinadas minimizando a divergência de Kullback-Leibler [43] para a distribuição conjunta P e Q conforme a Eq. (2.11):

$$C(Y) = KL(P||Q) = \sum_{i\neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2.11}$$

Segue abaixo uma versão simplificada do algoritmo [44]:

Um exemplo de redução de dimensionalidade utilizando o t-SNE é mostrado na Figura 2.6, em que o conjunto de dados *Iris* de 4 dimensões é representado em um gráfico bidimensional. Ambos os gráficos do PCA e t-SNE separam bem o conjunto *Iris*, porém, percebe-se que a projeção bidimensional realizada pelo t-SNE aglomera os pontos em grupos menos espaçados.

Algorithm 1: Versão simplificada do TSNE

Input: Conjunto multidimensional *X*,
Perplexidade da função do custo *p*,
número de iterações *T*,
taxa de aprendizado μ,
momento $\alpha(t)$
Output: Conjunto de dimensão reduzida $Y^{(T)} = \{y_1, y_2, ..., y_n\}$
Compute os pares de afinidade $P_{i|j}$ com perplexidade *p* utilizando a Eq. (2.8);
Compute P_{ij} utilizando a Eq. (2.9);
Inicialize $Y^{(0)}$ aleatoriamente;
faça $t = 1$;
while $t < T$ **do**
    Compute as afinidades q_{ij} utilizando a Eq. (2.10);
    Compute a divergência de Kullback-Leibler C utilizando a Eq. (2.11);
    faça $Y^{(t)} = Y^{t-1} + \mu C + \alpha(Y^{t-1} - Y^{t-2})$;

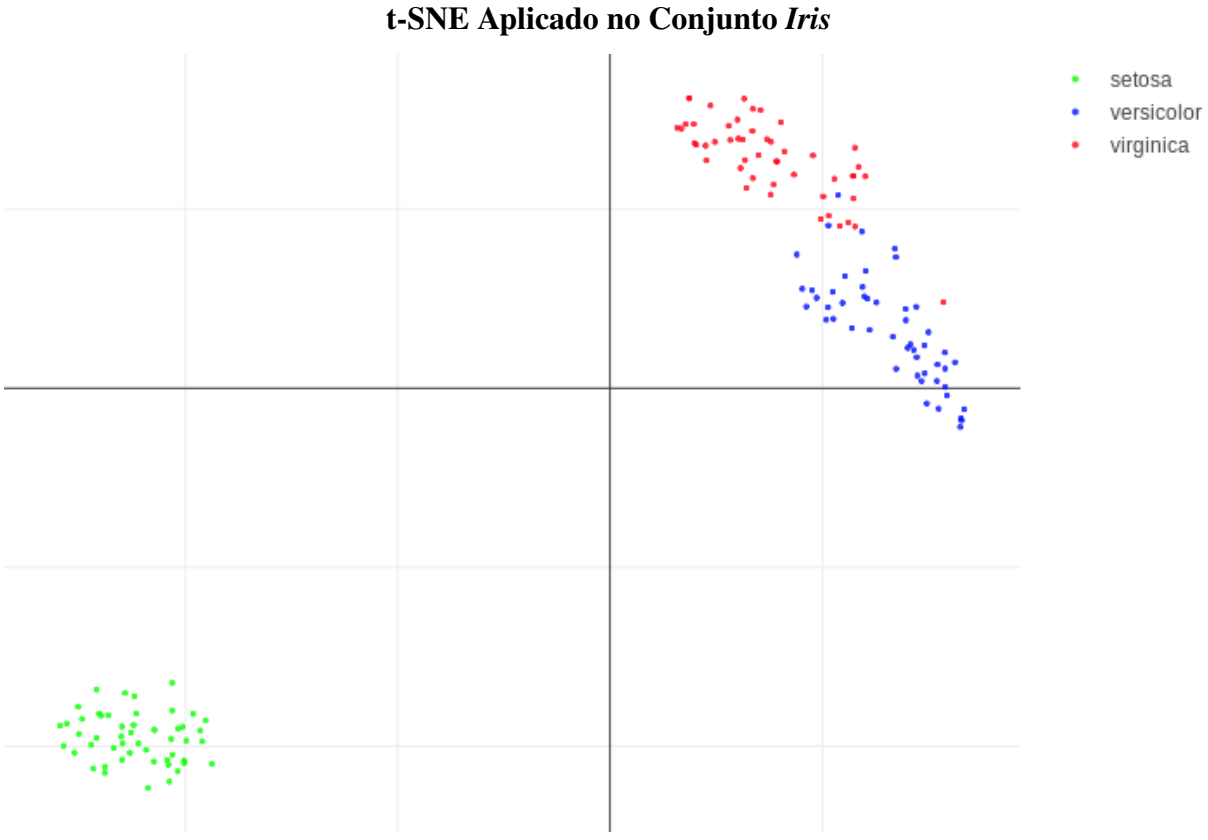


Figura 2.6: Exemplo de aplicação do t-SNE.

2.4 Aprendizado de Máquina

Aprendizado de máquina é definido como a capacidade de um computador se adaptar a novas circunstâncias, detectar e explorar padrões [45]. Técnicas de aprendizado de máquina são utilizadas em diversas ocasiões como previsão de preços de produtos [46], detecção de imagens [47], carros autônomos [48], reconhecimento de voz [49], detecção de sentimentos [50], entre outros.

A aprendizagem pode ser realizada de forma supervisionada ou não supervisionada. Quando supervisionada, a técnica é ajustada a partir de um conjunto de pares de entrada e saída, assim, sendo possível mapear uma função a partir dos pares conhecidos. No caso não supervisionado o processo é realizado somente com o conjunto de entrada. Em ambos os casos, ao realizar predições, um problema de aprendizagem pode ser de classificação ou regressão. Quando o resultado é um conjunto finito de classes (eg. verdadeiro ou falso) está sendo realizada uma classificação e, regressão caso seja um conjunto contínuo (eg. temperatura do dia seguinte) [45].

No âmbito de estudos educacionais, técnicas de aprendizado de máquina são utilizadas para classificação, predição das informações, entre outros [51]. Assim, com o intuito de identificar padrões, neste trabalho foram utilizadas técnicas de aprendizado de máquina para realizar agrupamento de dados por meio do algoritmo *k-means* e predições com o *K-NN* e o *Gradient Boosting*.

2.4.1 *K*-means

O *k*-means, classificado como um método de agrupamento de dados particional, é um dos algoritmos de agrupamento mais utilizados [52]. A partir de um conjunto de objetos que contenham somente valores numéricos $\boldsymbol{X}$. Sendo n e m o tamanho e dimensionalidade de $\boldsymbol{X}$ respectivamente e $k \leq n$ um número inteiro. O algoritmo *k*-means separa o conjunto $\boldsymbol{X}$ em k *clusters* tal que se minimize a soma dos erros quadráticos de cada grupo. Este processo pode ser formulado matematicamente como a minimização da função P descrita na Eq. (2.12) [53, 54]:

$$P(W, Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{i,l} d(x_i, q_l) \tag{2.12}$$

onde:

$$\sum_{l=1}^{k} w_{i,l} = 1, \quad 1 \leq i \leq n$$
$$w_{i,l} \in \{0, 1\}, \quad 1 \leq i \leq n \text{ e } 1 \leq l \leq k$$

W é uma matriz $n \times k$, $\boldsymbol{Q} = \{q_1, q_2, ..., q_k\}$ é um conjunto de objetos de mesmo domínio, e $d(\cdot, \cdot)$ é a distância Euclidiana definida na Eq. (2.1). Segue a abaixo uma versão simplificada do algoritmo [55]:

Algorithm 2: Versão simplificada do *k*-means

Input: Número de clusters *k*,
Conjunto de dados *D*,
Número máximo de iterações *T*
Output: Um conjunto com *k* clusters
Escolha *k* objetos de *D* arbitrariamente como centros iniciais dos clusters;
faça $t = 0$;
while $t < T$ **do**
    Realoque cada objeto de *D* para o cluster mais semelhante baseado no valor médio dos objetos do cluster;
    Recalcule os clusters para a média dos objetos presentes em cada cluster;
    Termine caso nenhuma mudança ocorra;

O algoritmo *k*-means dispõe de propriedades tais como: eficiência para o processamento de grandes conjuntos de dados [56]; somente funciona com conjuntos numéricos [56]; e, frequentemente finaliza o algoritmo retornando um resultado ótimo local [53].

2.4.2 K-Nearest Neighbor (K-NN)

O algoritmo de vizinhos mais próximos é um dos procedimentos de classificação não paramétrico mais simples [57]. Dado um conjunto de pares $Q = \{q_1, q_2, ..., q_n\}$ onde $q_i = (x_i, \theta_i)$ sendo x_i um ponto pertencente a um conjunto de dados e θ_i a categoria pertencente a um conjunto finito $M = \{1, 2, ..., m\}$ na qual x_i é classificado. Para classificar um novo ponto arbitrário x_j, primeiramente identificamos os seus *K*-vizinhos mais próximos, ou *Nearest Neighbors* seguindo a Eq. (2.13). Para o caso particular 1-NN, a classificação θ do ponto x_j se dá pelo mesmo θ do vizinho mais próximo encontrado.

$$min \quad d(x_i, x_j) = d(x_i, x_j), i = 1, 2, ...n \tag{2.13}$$

em que d é uma função de dissimilaridade [57].

Para o caso geral, **K-NN**, utiliza-se da Eq. (2.13) para selecionar os K vizinhos mais próximos. O ponto x receberá a classificação θ mais recorrente entre os vizinhos selecionados.

2.4.3 Gradient Boosting

O *gradient boosting* constrói um modelo de regressão ajustando sequencialmente uma função de aprendizado simples, em que o método dos mínimos quadrados é utilizado a cada iteração. Ao final do algoritmo, é gerada a função de estimação [58].

Para realizar estimações de funções é necessário uma variável de "saída" (*output*) ***Y*** e um conjunto de variáveis de "entrada" (*input*) ***X***. A partir de um conjunto de treino de treino $T = \{y_i, x_i\}_1^N$, tenta-se encontrar uma função $F^*(x) = Y$, em que, considerando todos os valores de ***T*** a função de perda $L(y, F(x))$ é minimizada para um valor esperado $E(X, Y)$ [59].

$$F^*(x) = arg \min_{F(x)} \; E(X, Y)L(y, (F(x))) \tag{2.14}$$

O procedimento de *boosting* aproxima a Eq. (2.14) por um somatório de acordo com a Eq. (2.15), onde $h(x; a)$ é uma função de aprendizado com parâmetros $a = \{a_1, a_2, ...\}$ e β_m é o coeficiente de expansão.

$$F(x) = \sum_{m=0}^{M} \beta_m h(x; a_m) \tag{2.15}$$

Os coeficientes de expansão $\{\beta_m\}_0^M$ e os parâmetros $\{a_m\}_0^M$ são ajustados de acordo com os conjuntos de treino seguindo a Eq. (2.16) a partir de um $F_0(x)$ aleatório.

$$(\beta_m, a_m) = arg \min_{\beta,a} \sum_{i-1}^{N} L(y_i, F_{m-1}(x_i)) + \beta h(x_i; a) \tag{2.16}$$

obtendo a $F_m(x)$ a partir Eq. (2.17).

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m) \tag{2.17}$$

O *gradient boosting* é uma aplicação do *boosting* em que $h(x; a)$ é ajustada pelo método de mínimos quadrados definindo a_m [58] conforme a Eq. (2.18):

$$a_m = arg \min_{\rho,a} \sum_{i-1}^{N} [y'_{im} - \rho h(x_i; a)]^2 \tag{2.18}$$

em que y'_{im} é o "pseudo"-resíduo definido na Eq. (2.19):

$$y'_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \tag{2.19}$$

Uma vez obtido a_m, calcula-se β_m conforme a Eq. (2.20):

$$\beta_m = arg \min_{\beta} \sum_{i-1}^{N} L(y_i, F_{m-1}(x_i) + \beta h(x_i; a_m)) \tag{2.20}$$

Assim, é possível simplificar a Eq. (2.16) por um processo mais simples de se calcular que utiliza o método de mínimos quadrados seguido por uma otimização de um único parâmetro de uma dada função de perda L [58]. Segue abaixo uma versão simplificada do algoritmo *gradient boosting* [58].

Algorithm 3: Versão simplificada do *gradient boosting*

Input: Conjunto de treinamento *T*,
Função de perda $L(y, F(x))$,
Número máximo de iterações *M*
Output: Função de regressão $F_M(x)$
Inicialize $F_0(x)$ utilizando a Eq. (2.14);
Faça $m = 1$;
while $m < M$ **do**
    Compute os pseudo-resíduos y'_{im} utilizando a Eq. (2.19);
    Ajuste a função de aprendizagem $h(x; a)$ para os pseudo-resíduos;
    Compute o coeficiente de expansão β_m utilizando a Eq. (2.20);
    Atualize a função de regressão $F_m(x)$ utilizando a Eq. (2.17);

2.5 Considerações Finais

As definições apresentadas neste capítulo foram selecionadas com base no que será desenvolvido no decorrer deste trabalho. Assim, definições como aprendizado de máquina e redução de dimensionalidades possibilitaram a compreensão dos dados educacionais, criação de modelos preditivos e o desenvolvimento de uma ferramenta interativa de visualização. Estes processos são descritos com mais detalhes nos Capítulos 3 e 4.

Capítulo 3

Visualização exploratória de Alunos Cotistas e Previsão de Desempenho

Em 2004, a Universidade de Brasília implementou um sistema de cotas raciais no processo de ingresso na instituição [60]. Tendo se passado mais de 10 anos, é natural que esse conjunto de alunos tenha gerado uma quantidade de dados educacionais suficientes para serem analisados e comparados com o conjunto de alunos não cotistas. Com o intuito de facilitar essa análise, realizou-se um estudo focado na visualização dos dados desses estudantes, provendo um maior entendimento da situação das cotas na universidade.

Em conjunto com a análise visual, foi desenvolvido um modelo de previsão de desempenho dos alunos do departamento de Ciência da Computação com o intuito de identificar se o estudante concluirá ou não o curso com base em seu desempenho acadêmico.

Para expor os dois tópicos, o resto deste capítulo foi dividido em 4 partes: Na Seção 3.1 são apresentados trabalhos previamente realizados sobre alunos beneficiados de ações afirmativas e previsão de desempenho de estudantes; na Seção 3.2 é detalhado o conjunto de dados, o pré-processamento para torná-lo utilizável e a metodologia utilizada; na Seção 3.3 expõe os resultados da visualização e do modelo de previsão; por fim, as considerações finais sobre o trabalho são discutidas na Seção 3.4.

3.1 Revisão de Literatura

Com o objetivo de entender o que vem sendo pesquisado na área de estudos de dados educacionais, bem como encontrar exemplos de técnicas para auxiliar este trabalho, procurou-se pesquisas com temas em predição de desempenho de estudantes e de análise de dados de alunos cotistas.

Saa et al. [61] apresenta um modelo de previsão de performance para estudantes no ensino superior. Com uma base de dados formada por informações de 270 estudantes contendo 21

atributos, os alunos foram separados em 4 classes referentes ao seu rendimento acadêmico (Excelente, muito bom, bom e aceitável), e, utilizando os softwares de aprendizado de máquina *Weka* e *RapidMiner* aplicou-se quatro variações (C4.5, ID3, CART e CHAID) do algoritmo de árvore de decisão e o Naive Bayes para fazer a classificação. A pesquisa atingiu uma taxa de 40% de acurácia com a árvore de decisão CHAID, um rendimento melhor comparado com os outros quatro métodos que obtiveram acurácias entre 33% e 36%. O estudo conclui que as notas dos alunos não depende somente do seus esforços acadêmicos, tendo outros fatores exercendo uma influência igual ou até mesmo maior.

Fernandes et al. [62] utiliza dados de alunos do ensino médio de escolas públicas do Distrito Federal para tentar encontrar alunos com maior chance de reprovação considerando os dois primeiros meses do ano letivo. Foram analisadas informações de 485872 estudantes, sendo 238575 do ano de 2015 e 247297 de 2016. Para fazer a predição, foram criados 2 modelos: um utilizando somente os dados prévios ao início do ano letivo (DT1); outro considerando também informações escolares referentes aos dois primeiros meses de aula (DT2).

O estudo extrai informações e implementa as predições utilizando o algoritmo *Gradient Boosting*, onde se obteve até $0,96$ na curva de Característica de Operação do Receptor (COR) para o conjunto de validação de DT1 e $0,91$ para o DT2. Além disso, é mostrado que para a classificação do DT1 as variáveis bairro do estudante, escola, idade e a cidade são os fatores mais relevantes para o algoritmo. Já considerando o DT2, ou seja, os dados após o início das aulas, os atributos nota, bairro do estudante, escola e matéria escolar são as mais influentes.

Costa et al. [63] desenvolve um modelo de predição de sucesso escolar dos alunos cotistas da Universidade Federal da Paraíba. No estudo, os alunos são considerados bem sucedidos quando apresentam um coeficiente escolar acima de 7. Com o auxílio da ferramenta *Weka*, foram utilizados 5 algoritmos de aprendizagem de máquina diferentes (J48, Naive Bayes, SMO, IBK e Multilayer Perceptron) tendo como entrada 4 atributos (Forma de ingresso, Tipo de cota, curso e nota de ingresso) pertencentes a uma base de dados de 10130 instâncias. O seu método de classificação obteve um melhor resultado com o algoritmo IBK, registrando uma acurácia média de $88,7\%$.

Dario et al. [64] implementa um estudo de caso sobre os alunos ingressantes no campus de Florianópolis da Universidade Federal de Santa Catarina entre os anos de 2013 a 2016 comparando o rendimento acadêmico e evasão entre os alunos cotistas e não cotistas. Foram utilizados dados de 18015 estudantes de 82 cursos diferentes analisando a forma de ingresso, a situação acadêmica e a média geral dos estudantes. Destes, $32,8\%$ ingressaram por meio do sistema de cotas e $67,2\%$ em ampla concorrência.

Como resultado, o estudo constatou diferenças tanto no rendimento acadêmico quanto na taxa de evasão. Os estudantes cotistas obtiveram como nota média $6,06$ e uma taxa de evasão de $12,81\%$, já os ingressantes de ampla concorrência alcançaram $6,53$ de média e $16,25\%$ de

evasão. O estudo conclui enfatizando que as diferenças no desempenho entre os grupos variam de acordo com o departamento e a área de estudo dos alunos. Cursos de ciências humanas e sociais não apresentam divergências significativas. No entanto, ciências exatas e da terra, como engenharias, apresentam maior discrepância quanto às notas.

Bonfim et al. [60] apresenta um estudo observacional dos alunos cotistas e não cotistas pertencentes a Universidade de Brasília que ingressaram no segundo semestre de 2004 e de 2009. Na análise, para encontrar fatores que influenciam no rendimento acadêmico e na evasão dos estudantes, foram utilizados dados sociodemográficos (idade, sexo, renda, escolaridade dos pais), pré-universitários (tipo de ensino médio, se já trabalhou, se já foi universitário) e de ingresso na universidade (nota do vestibular e curso).

Quanto ao rendimento acadêmico, nota-se uma diminuição da diferença entre cotistas e não cotistas comparando alunos de 2004 com aqueles de 2009. Além disso, percebe-se uma tendência de um melhor desempenho dos alunos que: não cursaram algum curso em universidade anteriormente; os pais possuem uma maior escolaridade; ou não pertencem a cursos de exatas. Considerando evasão nota-se uma maior propensão à evasão os alunos que: são mais velhos; já cursaram alguma universidade; já trabalharam; ou fizeram supletivo.

Os estudos apresentados indicam que a procura de formas de prever a evasão de estudantes é um objetivo comumente pesquisado, além disso, percebe-se que existe um interesse em comparar os alunos cotistas e não cotistas. Dessa forma, vemos a relevância dos temas abordados nesse trabalho, em que é realizado a predição de desempenho e a comparação entre os alunos com foco em técnicas de visualização de dados.

3.2 Metodologia Proposta

Para auxiliar o desenvolvimento do estudo e melhorar o resultados obtidos, definiu-se uma metodologia contendo cinco etapas sequenciais representadas na Figura 3.1. As cinco etapas englobam o processo da pesquisa referente a visualização exploratória dos dados e a predição de desempenho dos estudantes, que são descritas da seguinte forma:

1. **Entendimento do problema e dos dados**: Fundamental para evitar a realização de trabalhos desnecessários e otimizar os resultados da pesquisa, a etapa inicial consiste em entender os dados e definir os objetivos possíveis com o conjunto de dados disponíveis. Após a compreensão dos dados, descritos na Seção 3.2.1, foram definidos dois objetivos principais: fazer uma comparação entre os alunos cotistas e não-cotistas tentando indicar, caso exista, possíveis motivos para as diferenças encontradas; e criar um modelo preditivo para os estudantes tentando identificar se eles concluirão ou não o curso, disponibilizando ao departamento mais informações para auxiliar os alunos nos estudos e evitar possíveis evasões.

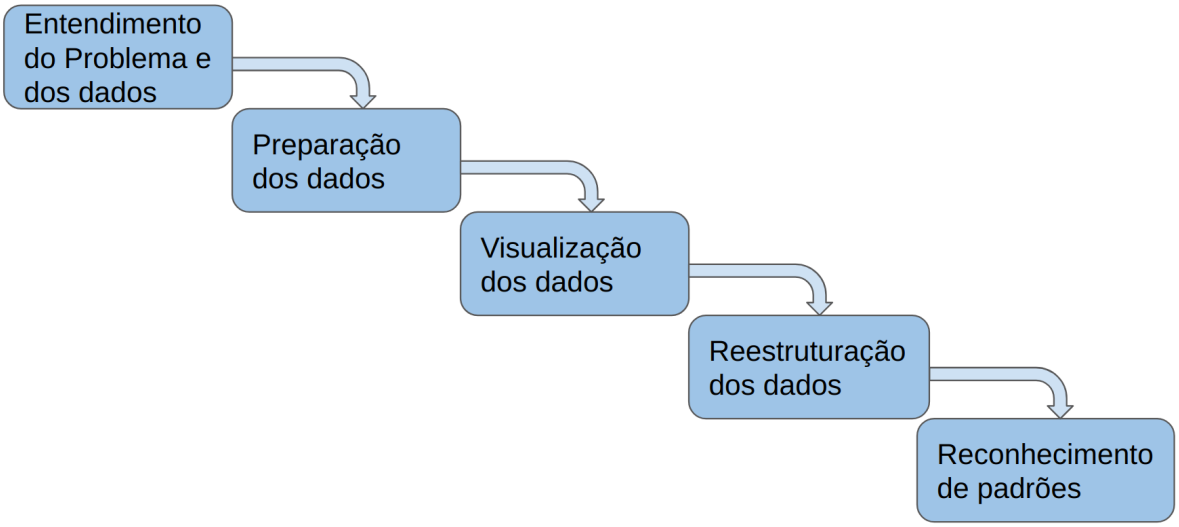


Figura 3.1: Metodologia criada para a análise visual dos dados e predição de desempenho.

2. **Preparação dos dados**: Após a definição dos objetivos da pesquisa e compreensão dos dados disponíveis, realizou-se o pré-processamento dos dados para que o conjunto esteja em um formato compatível com a implementações de algoritmos de visualização e redução de dimensionalidade. Assim, o conjunto de dados foi ajustado de forma a facilitar sua visualização e o seu processamento. O processo, mais detalhado na Seção 3.2.2, envolveu descartar variáveis redundantes para o estudo, transformar tipos de variáveis e a criação de conjuntos de dados derivados do conjunto original.

3. **Visualização dos dados**: Como método de extração de informações para comparação entre os alunos, escolheu-se a visualização exploratória dos dados. Na tentativa de encontrar padrões e correlações, foram aplicadas técnicas de redução de dimensionalidade e gerados gráficos como *Heat maps* e coordenadas paralelas. Juntamente com a geração dos gráficos, são analisadas e extraídas informações dos dados utilizados. A partir dessas informações, são criadas conclusões e hipóteses para as situações encontradas na visualização. Este processo, descrito com mais informações na Seção 3.3.1, visa realizar a extração de informações dos dados educacionais de acordo com a necessidade do usuário além de torná-las mais intuitivas.

4. **Reestruturação dos dados**: A partir das informações adquiridas na comparação dos alunos, modificou-se a base de dados para obter melhores resultados na aplicação de métodos de aprendizado de máquina. As mudanças nos dados são descritas na Seção 3.3.2.

5. **Reconhecimento de padrões**: Com os dados preparados, aplicou-se os algoritmos de

classificação *K-NN* e o *Gradient Boosting*, visando predizer se os alunos abandonarão ou não seus cursos levando em consideração as notas obtidas nos primeiros semestres, tornando assim possível a instituição auxiliar os estudantes com maior chance de abandono de uma forma mais específica. Mais informações sobre o processo são descritas na Seção 3.3.2.

3.2.1 Conjunto de Dados

Na realização deste estudo, foram utilizados dados extraídos a partir de históricos escolares dos alunos pertencentes ao departamento de Ciência da Computação da UnB obtidos através do Sistemas Acadêmicos da Universidade de Brasília (SIGRA). Assim, o conjunto de dados abrange os estudantes dos seguintes cursos: Computação; Engenharia de Redes de Comunicação; Ciência da Computação; Engenharia Mecatrônica; Engenharia de Computação; Engenharia de Software; Informática.

Inicialmente, os dados eram formados por 281.024 instâncias na qual cada uma representa uma matéria cursada possuindo os seguintes 26 atributos:

- **ID Aluno:** número de identificação de cada aluno;
- **Sexo:** o gênero de cada aluno, tendo dois possíveis valores: *masculino ou feminino*;
- **Data de Nascimento:** data de nascimento dos estudantes seguindo o formato AAAA/MM/DD;
- **UF de Nascimento:** sigla de duas letras referentes a Unidade da Federação que o aluno nasceu;
- **Cotista:** tipo de ingresso na universidade, tendo dois possíveis valores: *Cotistas ou Não Cotistas*;
- **Tipo de Escola:** tipo de escola em que o aluno estudou antes de ingressar na universidade, com três categorias: *escola pública, escola particular e não informado*;
- **Raça:** Raça declarada pelo estudante ao ingressar na universidade, podendo assumir os valores *preta, parda, branca, amarela, não cadastrada, não informada ou não dispõe de informação*;
- **Curso:** o curso na qual o estudante pertence podendo ser *Computação, Engenharia de Redes de Comunicação, Ciência da Computação, Engenharia Mecatrônica, Engenharia de Computação, Engenharia de Software ou Informática*;
- **Opção:** a opção escolhida pelo aluno ao ingressar na universidade podendo ser *Computação, Engenharia de Redes de Comunicação, Ciência da Computação, Engenharia de*

Controle e Automação, Engenharia de Computação, Engenharia de Software ou Informática;

- **Período de Ingresso na Unb:** o período na qual o aluno ingressou na universidade com valores entre o primeiro semestre de 1991 ao segundo semestre de 2016;
- **Período de ingresso na opção:** o período na qual o aluno ingressou na opção com valores entre o primeiro semestre de 1991 ao segundo semestre de 2016;
- **Forma de Ingresso na UnB:** a forma como o aluno ingressou na universidade, podendo ser *Vestibular, Matrícula Cortesia, Transferência Obrigatória, Programa de Avaliação Seriada (PAS), Transferência Facultativa, Convênio Andifes, Acordo Cultural (PEC-G), Convênio Internacional, PEC-Peppfol Graduação, Vestibular para mesmo Curso, Portador Diploma Curso Superior, Enem, Sistema de Seleção Unificada (Sisu) ou Refugiado*;
- **Período de saída da opção:** o período em que o aluno saiu da opção, com valores entre o primeiro semestre de 1992 ao começo de 2017;
- **Forma de saída da opção:** a forma como o aluno saiu da opção, podendo ser *Novo Vestibular, Desligamento Abandono, Formatura, Desligamento por não cumprir condição ,Desligamento Voluntário , Mudança de Curso, Repetir 3 vezes na mesma disciplina obrigatória, Desligamento por Decisão Judicial, Transferência, Falecimento, Desligamento por Força de Convênio, Ativo, Vestibular para outra Habilitação, Desligamento por Jubilamento, Mudança de Turno, Desligamento por Falta de Documentação, Desligamento por Força de Intercâmbio ou Outros*;
- **Ano e Semestre:** período na qual a matéria foi cursada pelo estudante;
- **Média do período:** a média das notas do aluno no semestre;
- **Mín. Créd. Formatura:** quantidade mínima de créditos necessários para a formatura;
- **Créditos cursados no total:** quantidade de créditos já cursados pelo aluno;
- **Créditos integralizados no total:** quantidade de créditos já integralizados pelo estudante;
- **Créditos a integralizar no total:** créditos a serem integralizados para o estudante poder formar;
- **Créditos cursados no semestre (com aprovação):** quantidade de créditos obtidos pelo aluno no semestre que cursou a matéria;
- **Código da disciplina:** o código da disciplina registrado no matricula web;

- **Nome da disciplina:** o nome da disciplina referente ao seu código registrado no matricula web;
- **Créditos disciplina:** a quantidade de créditos da disciplina;
- **Menção na disciplina:** a menção obtida pelo aluno ao cursar a disciplina, podendo ser classificada como *MM, SS, MS, MI, II, TR, SR, CC, AP, TJ, DP*.

3.2.2 Preparação dos Dados

Após a identificação dos dados à disposição, para possibilitar a implementação de algoritmos de visualização e a análise dos dados, realizou-se uma busca por inconsistências na qual verificou-se que não haviam anomalias (e.g. disciplina com quantidade de créditos negativos), ruídos (e.g. data de nascimento infactível) e nem redundâncias (e.g. instâncias repetidas). Em seguida, para formar o primeiro conjunto de dados utilizável (CD1), foram descartados os atributos *UF de Nascimento, Opção, Forma de Ingresso na UnB, Período de ingresso na opção, Nome da disciplina, Mín. Créd. Formatura, Créditos a integralizar no total e Créditos integralizados no total* por serem considerados irrelevantes para a análise ou cujo valor já estavam sendo representados por alguma outra variável. Além disso, realizou-se a adição dos atributos *Idade* e *Semestre* calculados através dos valores Ano e semestre, data de nascimento e Período de ingresso na UnB. Essas variáveis representam, respectivamente, a idade e o semestre do aluno quando a matéria foi cursada.

Selecionados os atributos a serem utilizados, para facilitar a implementação de métodos de redução de dimensionalidade e a representação gráfica, todos atributos categóricos foram substituídos por equivalentes numéricos, pois os métodos utilizados não suportam dados textuais na execução do seu algoritmo. A Tabela 3.1 mostra as conversões realizadas no processo, em que, com exceção dos atributos Curso e Raça, os valores foram escolhidos para que seja possível calcular a similaridade entre os valores das variáveis. Por fim, a variável *Forma de saída da opção* foi substituída por um novo atributo chamado *Estado da Graduação* em que define-se, de forma numérica, se o aluno está *cursando*, *abandonou* ou *formou* no curso.

A partir do CD1 criou-se um novo conjunto (CD2), em que, cada instância representa um dos 7683 alunos e, seus atributos, apresentam a quantidade de vezes que o estudante cursou cada uma das disciplinas (já cursada por algum estudante do departamento), totalizando 1959 matérias diferentes. Além disso, para acrescentar as informações sobre o estudante, adicionou-se os atributos *Curso, Estado da Graduação, Raça, Sexo, Cotista e ingresso na UnB*.

Atributo	**Transformação**
Curso	Computação Licenciatura $\rightarrow 0$ Engenharia de Redes de Comunicação $\rightarrow 1$ Ciência da Computação bacharelado $\rightarrow 2$ Engenharia Mecatrônica $\rightarrow 3$ Engenharia de Computação $\rightarrow 4$ Engenharia de Software $\rightarrow 5$ Informática $\rightarrow 6$
Raça	Não informado $\rightarrow 0$ Não cadastrada $\rightarrow 0$ Não dispõe da informação $\rightarrow 0$ Parda $\rightarrow 1$ Amarela$\rightarrow 2$ Preta $\rightarrow 3$ Indígena $\rightarrow 4$ Branca $\rightarrow 5$
Tipo Escola	Não Informada $\rightarrow 0$ Particular $\rightarrow -1$ Pública $\rightarrow 1$
Estado da Graduação	Cursando $\rightarrow 0$ Abandonou $\rightarrow -1$ Formou$\rightarrow 1$
Sexo	M $\rightarrow 0$ F $\rightarrow 1$
Menção na disciplina	AP$\rightarrow 3$ CC$\rightarrow 3$ DP$\rightarrow 3$ II$\rightarrow 1$ MI$\rightarrow 2$ MM$\rightarrow 3$ MS$\rightarrow 4$ SS$\rightarrow 5$ SR$\rightarrow 0$ TJ$\rightarrow 0$ TR$\rightarrow 0$

Tabela 3.1: Tabela de conversões utilizada na preparação dos dados.

3.3 Resultados Experimentais

3.3.1 Visualização Exploratória

Para realizar uma comparação entre os alunos cotistas e não cotistas, aplicou-se técnicas de visualização de dados em CD1 e CD2. Os gráficos gerados foram utilizados para encontrar padrões, correlações e extrair as informações inicialmente almejadas na fase de entendimento do problema.

Inicialmente, com CD1, avaliou-se o desempenho dos estudantes por meio de suas taxas de aprovação. A Figura 3.2 apresenta um gráfico da porcentagem de aprovação nas disciplinas por semestre de todos os alunos do departamento. O gráfico mostra um maior rendimento nos semestres iniciais por parte dos alunos não cotistas, porém, a diferença tende a diminuir a cada período chegando ao ponto em que os cotistas reprovam menos em alguns dos semestres.

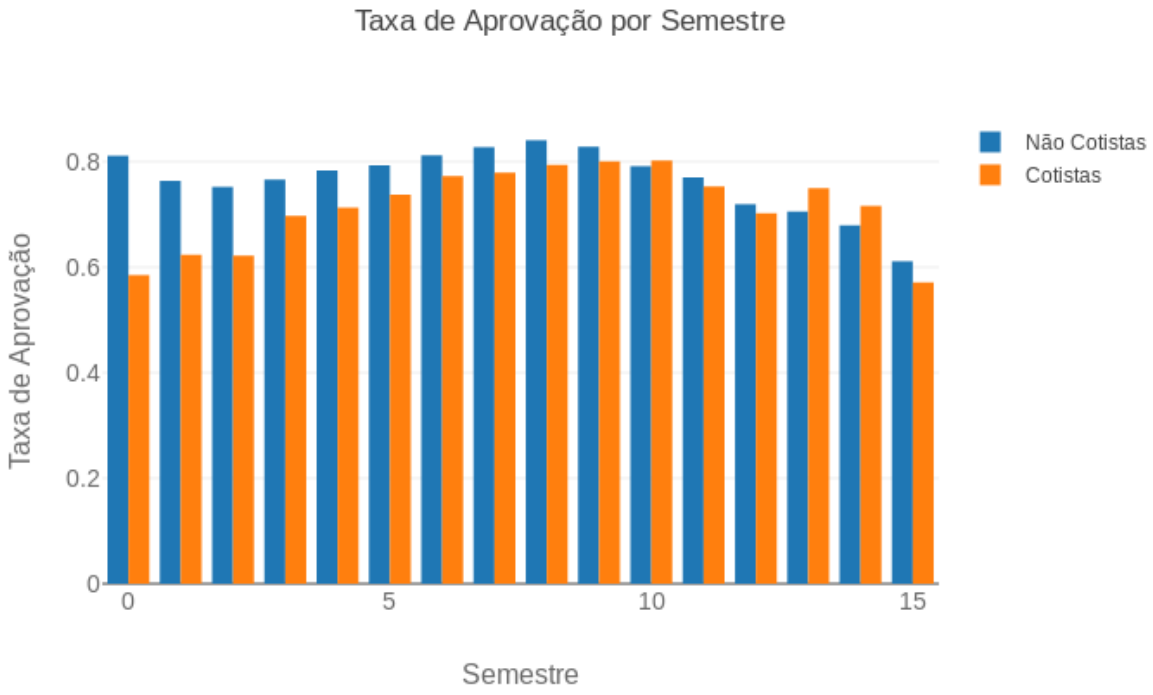


Figura 3.2: Gráfico de barras contendo a taxa de aprovação por semestre dos alunos cotistas e não cotistas.

Para compreender melhor a Figura 3.2, utilizando CD1, outro gráfico foi gerado focado nas taxas de evasão dos estudantes, ou seja, a porcentagem dos alunos que não formaram e não estão ativos no curso. Representado na Figura 3.3, o gráfico mostra a porcentagem de alunos que abandonaram o curso em seus respectivos semestres. Nota-se, dessa vez, um padrão diferente.

Nos primeiros semestres, as taxas são iguais, porém, no terceiro semestre, a porcentagem de abandono entre os alunos cotistas triplica, igualando-se ao outro grupo somente após a metade dos cursos.

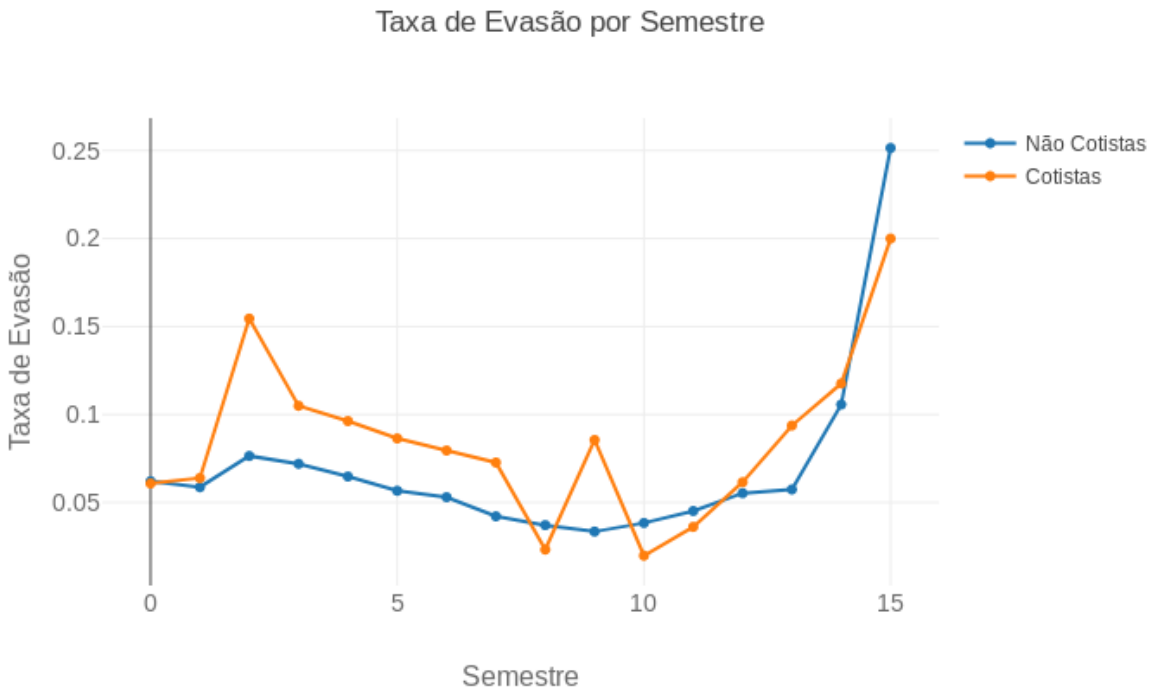


Figura 3.3: Gráfico de linha contendo a taxa de evasão por semestre dos alunos cotistas e não cotistas extraído a partir do CD1.

Observando as Figuras 3.2 e 3.3, percebe-se um comportamento diferenciado de todos os grupos de estudantes nos primeiros semestres. Nota-se, por exemplo, que após o pico de evasão do terceiro semestre, o desempenho dos alunos cotistas melhora consideravelmente. Além disso, considerando a utilização de porcentagens e eliminando os três primeiros semestres, percebe-se que os dados se comportam de forma semelhante.

Analisando o CD2, encontra-se o problema da alta dimensionalidade. Dado que o conjunto de dados possui 7683 instâncias e 1966 atributos, torna-se necessário a implementação de algum processo de redução de dimensionalidade como a transformação ou seleção de características para representá-los em uma forma gráfica. Assim, com o interesse de se extrair mais informações dos dados, escolheu-se aplicar ambas as formas de redução de dimensionalidade em análises separadas.

Como critério de seleção, escolheu-se as 10 matérias que possuíam a maior razão entre a quantidade de vezes que uma disciplina foi cursada e o total de alunos que a fizeram, ou seja, as

matérias que possuíam a maior média de vezes que um estudante cursa a disciplina. A Figura 3.4 mostra as disciplinas selecionadas juntamente com o valor de sua razão. Considerando a comparação entre cotistas e não cotistas, é interessante notar que, em geral, existe uma pequena diferença entre a média de vezes que os estudantes tenham que cursar as matérias para serem aprovados. A Tabela I.1 (no anexo) contém o nome das disciplinas referentes a cada código utilizado.

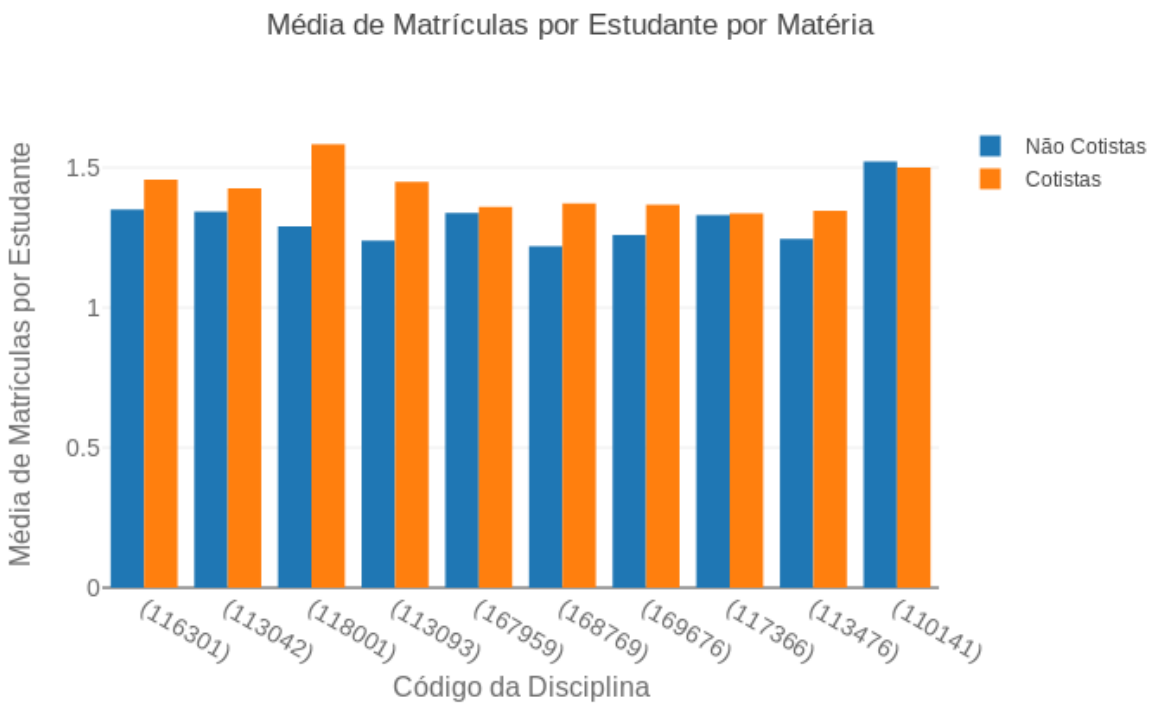


Figura 3.4: Gráfico comparativo entre alunos cotistas e não cotistas da taxa média de inscrição por disciplina extraído a partir do CD2.

Selecionadas as disciplinas a serem estudadas, com o intuito de identificar alguma relação com a evasão dos estudantes, aplicou-se a correlação entre elas e o estado de graduação dos estudantes. A Figura 3.5 mostra uma comparação das correlações encontradas entre os alunos cotistas e não cotistas. Nota-se uma baixa correlação entre a quantidade de vezes que o aluno cursa a disciplina e o Estado de Graduação do estudante, porém, encontra-se relações entre algumas matérias, como as de código 116101, 113042, 118001 e 113093 em ambos os subconjuntos, indicando que os alunos que fazem mais vezes alguma destas disciplinas tendem a fazer mais vezes as outras três. Também encontra-se correlações negativas, nota-se, por exemplo, que alunos não cotistas que fazem as disciplinas 117366 ou 113476 tendem a fazer com menos frequência as outras disciplinas selecionadas.

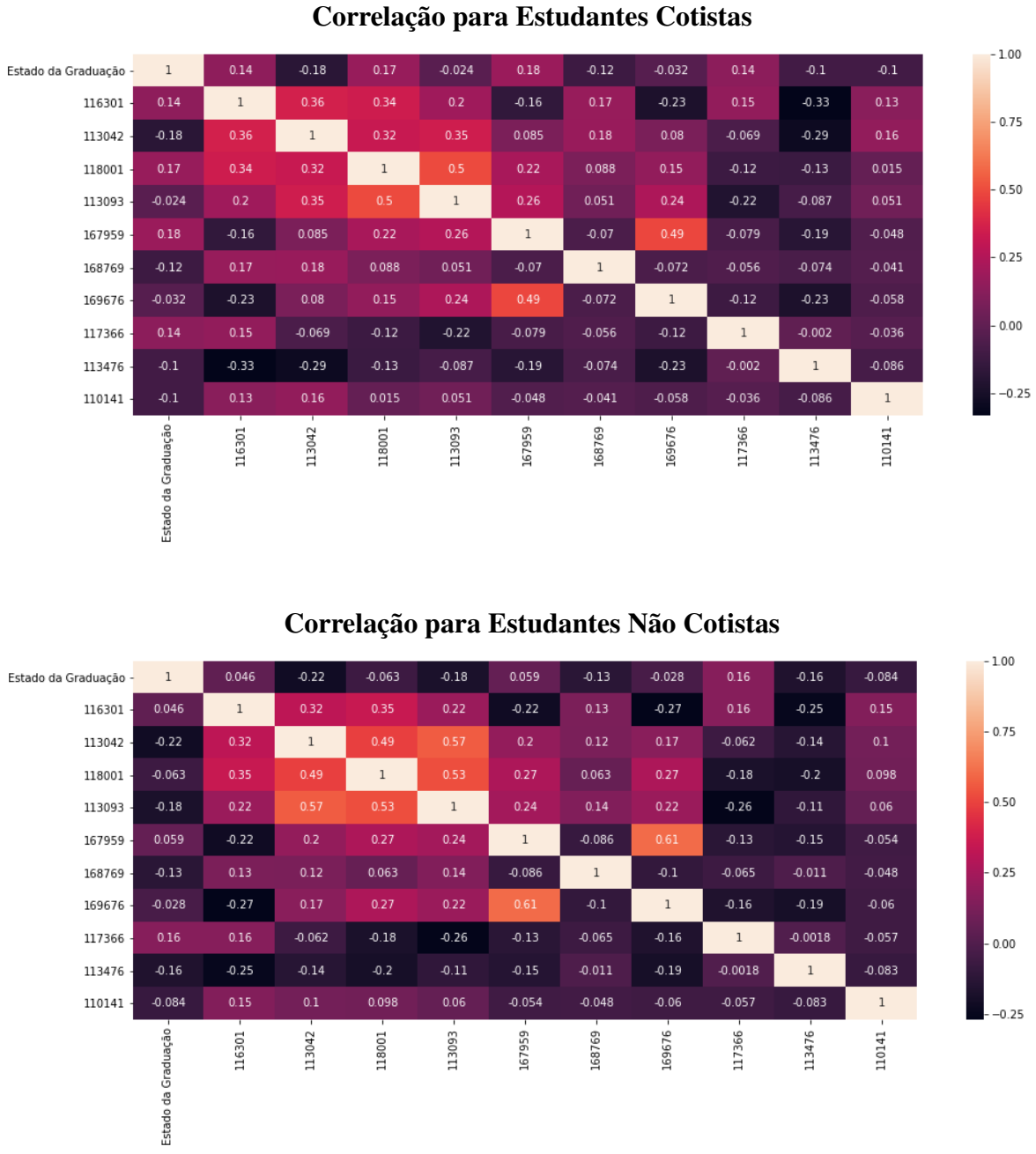


Figura 3.5: *Heat map* da correlação entre as disciplinas e o estado de graduação dos estudantes cotistas e não cotistas extraído a partir do CD2.

Em seguida, procurou-se padrões por meio da visualização baseada em coordenadas paralelas. A Figura 3.6 apresenta as disciplinas selecionadas juntamente com o curso, raça, sexo e o tipo de escola dos estudantes separados em cotistas e não cotistas, em que seus valores apresentados no *eixo-y* do layout estão descritos na Tabela 3.1. Pelos gráficos, nota-se uma tendência aleatória, impossibilitando a descoberta visual de padrões entre os grupos. Porém, novamente

as disciplinas 116101, 113042 e 118001 se destacam, sendo aquelas que ao menos um estudante cursou 8 ou mais vezes.

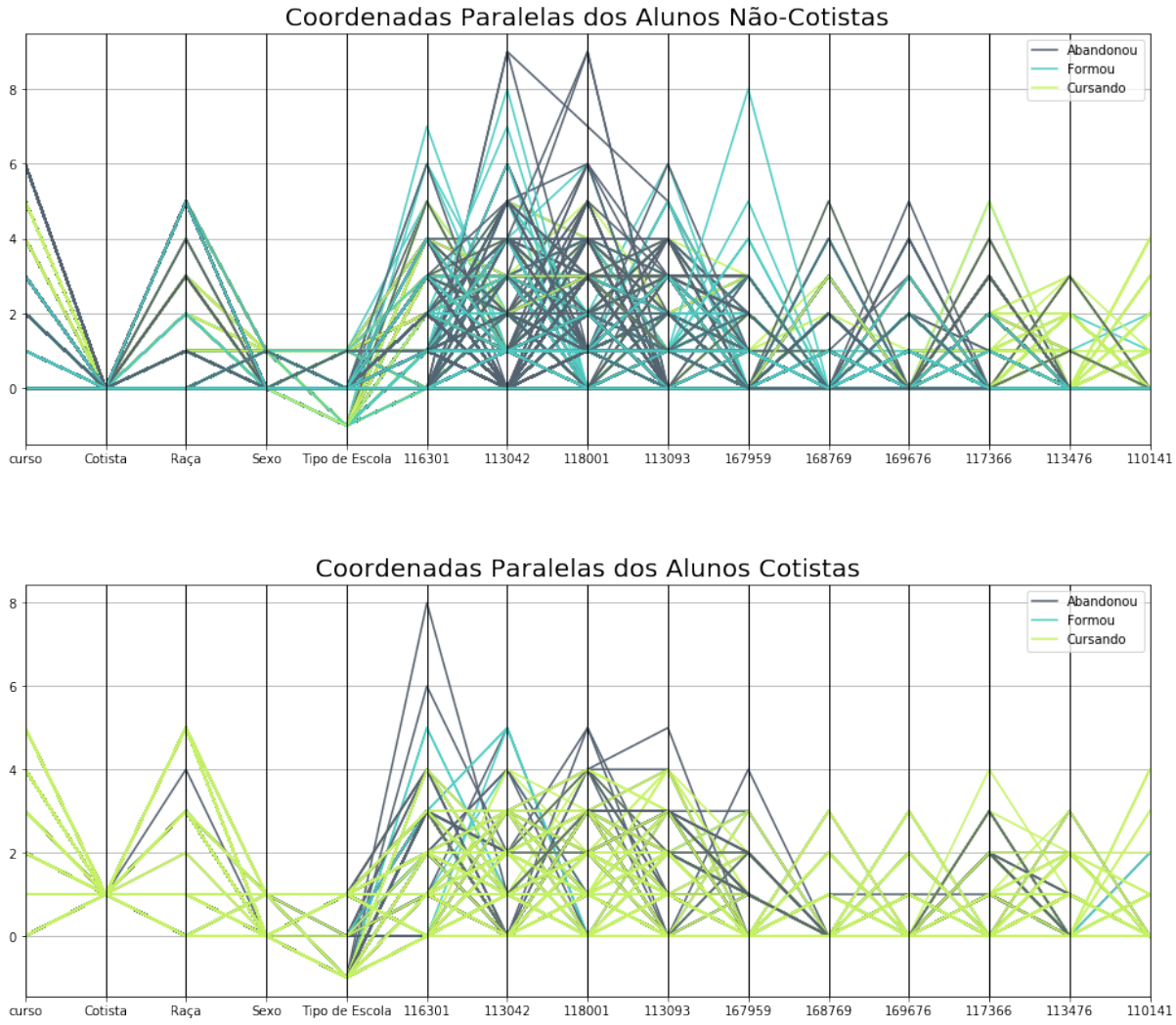


Figura 3.6: Gráficos de coordenadas paralelas para as disciplinas com maior taxa média de inscrição extraído a partir do CD2.

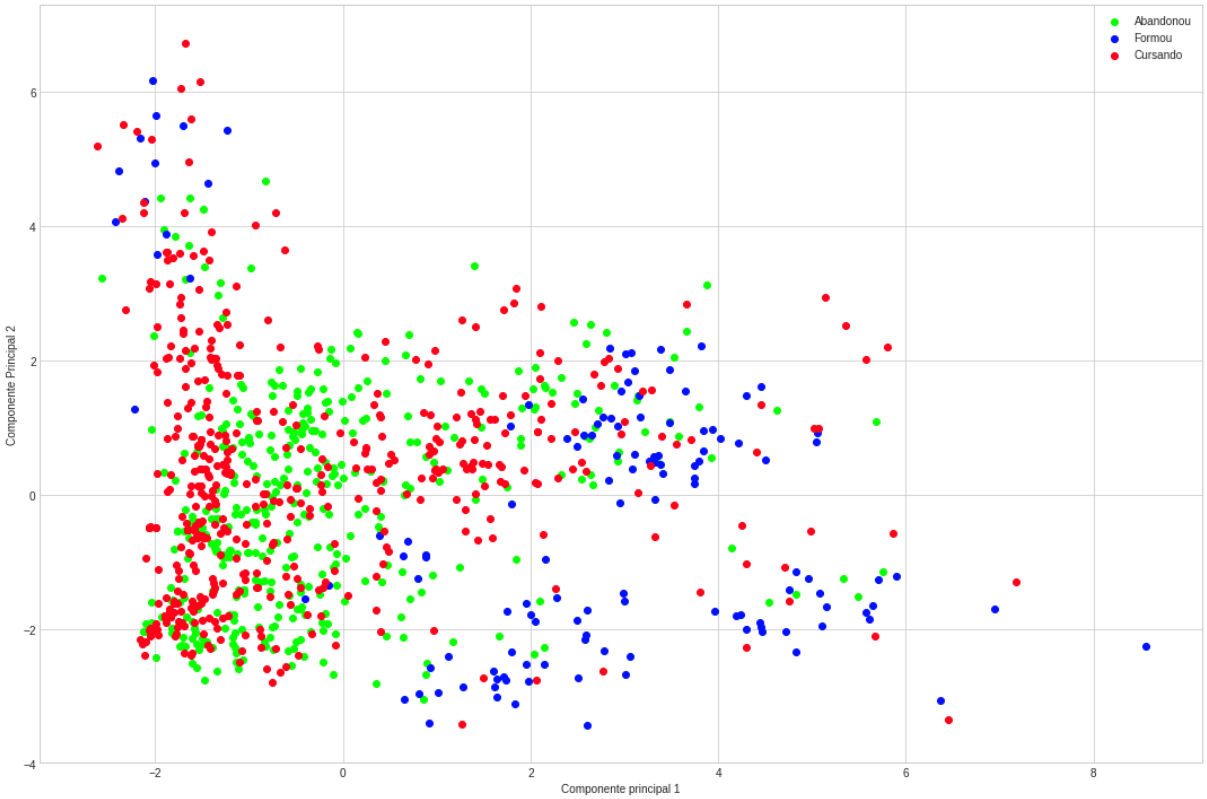
Todas as três matérias que se destacaram, tanto na correlação quanto nas coordenadas paralelas, eram disciplinas obrigatórias pertencentes aos primeiros semestres dos cursos estudados. Mostrando mais uma vez o padrão de desempenho identificado por meio da Figura 3.2, em que as matérias dos primeiros semestres apresentam uma maior taxa de reprovação comparadas com as outras.

Transformações de Características

Como mencionado anteriormente, para visualizar o CD2, utilizou-se técnicas de transformações de características na redução de dimensionalidade para a descoberta de padrões e verificar se os dados poderiam ser separados de acordo com o estado do estudante na universidade. Assim, novamente, separou-se o CD2 em dois subconjuntos de cotistas e não cotistas e aplicou-se duas técnicas, o PCA e o t-SNE.

A aplicação do PCA, exibida na Figura 3.7, apresenta os dois subconjuntos (alunos cotistas e não cotistas) de forma bidimensional. Nota-se, na visualização dos alunos não cotistas, que grande parte dos pontos estão sobrepostos, mesmo assim, identifica-se agrupamentos de estudantes que estão formados em partes específicas do *layout*. Este fato não ocorre na representação de alunos cotistas, em que os alunos estão mais dispersos e não percebe-se nenhum padrão característico.

PCA Aplicado no Subconjunto de Alunos Cotistas



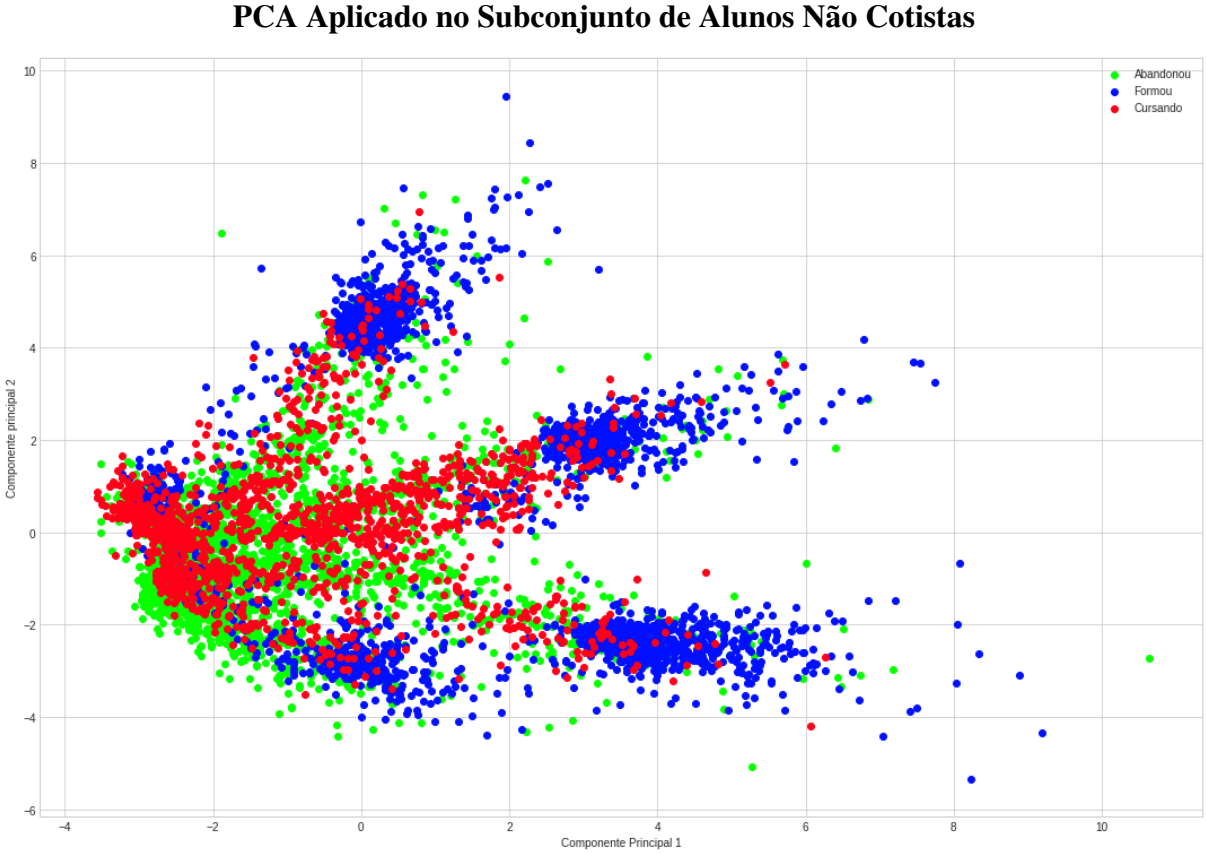


Figura 3.7: PCA aplicado no CD2 nos subconjuntos de alunos cotistas e não cotistas.

Com o intuito de encontrar uma representação que separa melhor as classes dos dois subconjuntos, aplicou-se o t-SNE em ambos. O algoritmo foi aplicado diversas vezes em cada subconjunto para encontrar a perplexidade do t-SNE que separou melhor as classes. Foram testadas perplexidades entre 10 e 50 sendo escolhida a perplexidade 20 que havia gerado a melhor visualização. A Figura 3.8 apresenta um *layout* bidimensional em que cada ponto representa uma instância do CD2 após a aplicação do t-SNE nos subconjuntos de estudantes cotistas e não cotistas. Diferentemente do PCA, percebe-se que as instâncias estão separadas de acordo com o *Estado do Curso*, onde os alunos que formaram aparecem mais distantes do centro, os que abandonaram tendem a ficar mais próximos do centro e os que ainda estão cursando aparecem entre ambos.

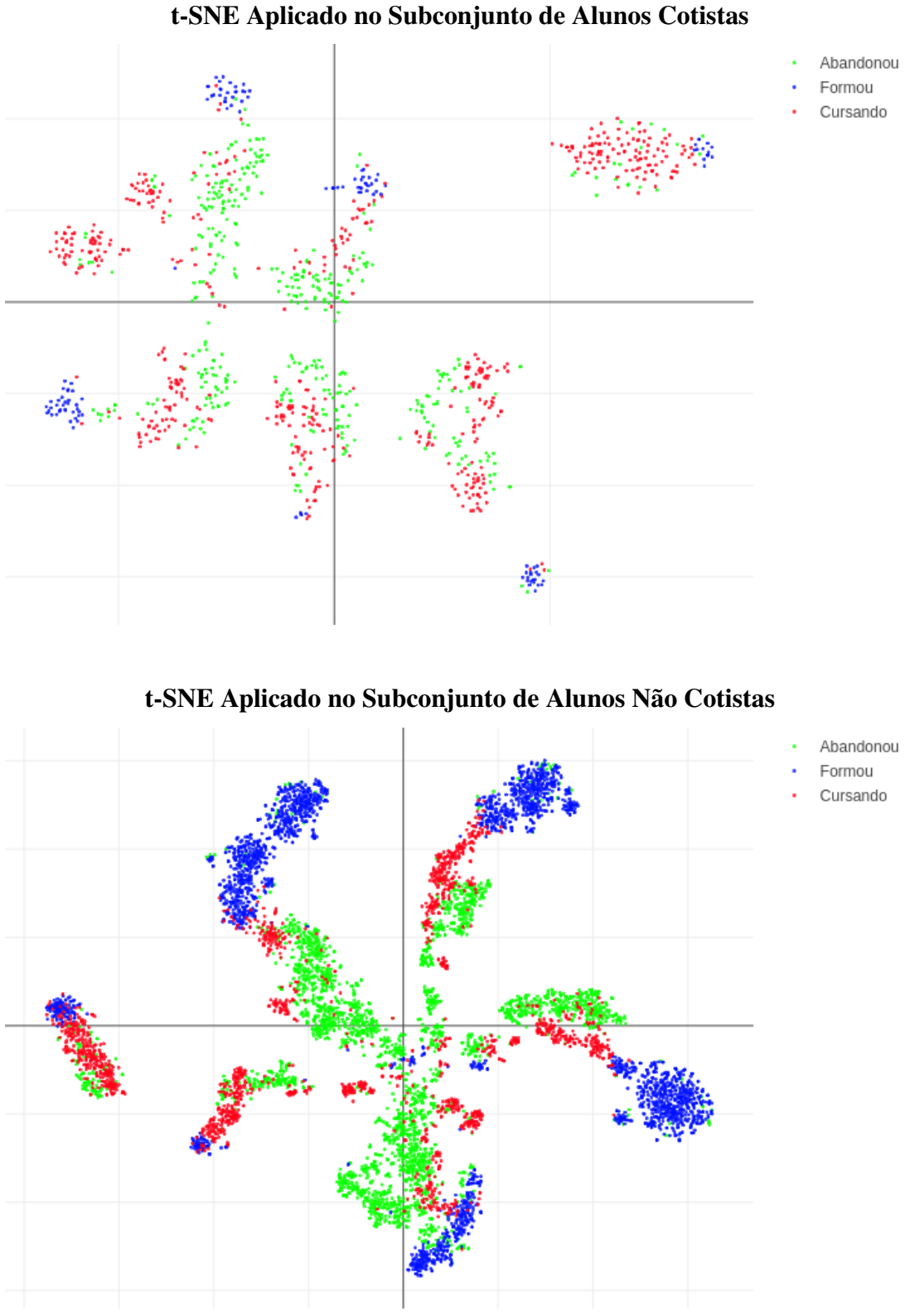


Figura 3.8: t-SNE aplicado no CD2 nos subconjuntos de alunos cotistas e não cotistas.

A melhor representação visual provida pelo algoritmo t-SNE é um indício de que os dados se comportam de forma não linear, fato anteriormente abordado em [65]. A representação dos

subconjuntos também gera indícios de que é possível criar um modelo classificatório com base no CD2, pois apresenta uma separação visível entre as classes dos estudantes. Além disso, nota-se a necessidade de uma ferramenta de seleção para saber quais são os estudantes pertencentes aos grupos apresentados nas Figuras 3.8 e 3.7 e, extrair mais informações sobre os alunos.

3.3.2 Previsão de Desempenho

Com o conhecimento sobre o conjunto de dados adquirido na Seção 3.3.1, implementou-se um modelo de previsão de desempenho dos alunos indicando uma probabilidade do estudante abandonar o curso. Para isso, criou-se um terceiro conjunto de dados (CD3). Gerado a partir do CD2, este conjunto possui os mesmos 1966 atributos, porém, ao invés de indicar somente a quantidade de vezes que o aluno cursou cada disciplina, cada variável representa a média das notas do estudante na matéria, acrescentando informações sobre o rendimento dos alunos no conjunto por ser um valor representativo do desempenho do aluno.

Com o objetivo de identificar os estudantes com maior risco de evasão e levando em consideração o comportamento diferenciado dos alunos nas matérias iniciais dos cursos, separou-se as disciplinas referentes aos fluxos dos três primeiros semestres de cada estudante. Porém, como cada graduação possui um fluxo diferente e os modelos de previsão necessitam de um conjunto de entrada de tamanho fixo, o CD3 foi fracionado entre os sete cursos listados na Seção 3.2.1.

Entre os sete cursos, notou-se que a universidade não disponibiliza mais a opção Informática, tornando desnecessário formular um modelo de classificação específico. Além disso, a graduação de Engenharia de Software também foi descartada por não apresentar um fluxo próprio nos primeiros 3 semestres. Assim, criou-se um total de 5 modelos referentes a *Computação, Engenharia de Redes de Comunicação, Ciência da Computação, Engenharia Mecatrônica e Engenharia de Computação*.

Na classificação, utilizou-se os algoritmos *K-NN* e o *Gradient Boosting* por serem comumente utilizados [66, 67, 68, 69, 62, 70]. Para validar o resultado aplicou-se a *k-fold cross validation* [71] dividindo o conjunto de dados em 5 partes. Além disso, para verificar o desempenho da predição, utilizou-se a métrica do *F1-score* definido como $F1 = 2(1/PR + 1/RE)$ em que PR representa a precisão e o RE o *recall* [72].

A Tabela 3.2 apresenta o resultado das classificações para os alunos não cotistas. Nota-se um desempenho entre 0,02 a 0,04 melhor no *Gradient Boosting* que também apresenta mais constância na classificação apresentando um desvio padrão menor. A classificação chegou a atingir a 0.94 referentes aos estudantes de Engenharia de Computação e 0.87 nos alunos de Ciência da Computação.

Curso	K-NN	K-NN σ	GB	GB σ
Computação	0.7719	0.05	0.8101	0.02
Engenharia de Redes	0.7371	0.04	0.7765	0.03
Ciência da Computação	0.8341	0.04	0.8700	0.02
Engenharia Mecatrônica	0.7289	0.05	0.7558	0.09
Engenharia de Computação	0.9196	0.07	0.9405	0.03

Tabela 3.2: Resultados dos modelos de classificação para os alunos não cotistas em que é representada a média do *F1-score* obtidos em cada predição.

Aplicando o mesmo modelo em alunos cotistas, obteve-se os resultados apresentados na Tabela 3.3. Dessa vez, o algoritmo K-NN apresenta resultados melhores em dois dos cinco cursos, além disso, de maneira geral, nota-se um maior desvio padrão nos resultados, indicando uma maior variação dos resultados nos testes de validação. Neste caso, tanto a maior taxa de acerto quanto o desvio padrão pode ser uma consequência do menor número de alunos cotistas já formados. Isto pode explicar resultados como os 0.97 dos alunos de Engenharia da Computação em que apenas 5 alunos cotistas já formaram. Mesmo assim, a classificação alcançou médias acima de 0.80 em todos os cursos.

Curso	K-NN	K-NN σ	GB	GB σ
Computação	0.8593	0.05	0.8376	0.08
Engenharia de Redes	0.8117	0.11	0.8193	0.11
Ciência da Computação	0.8872	0.08	0.9116	0.05
Engenharia Mecatrônica	0.8588	0.10	0.7793	0.08
Engenharia de Computação	0.9472	0.04	0.9711	0.03

Tabela 3.3: Resultados dos modelos de classificação para os alunos cotistas em que é representada a média do *F1-score* obtidos em cada predição..

3.4 Considerações Finais

Este capítulo apresentou duas abordagens para a descoberta de conhecimento em dados educacionais. A visualização exploratória mostrou-se eficiente para encontrar informações dificilmente percebidas ao visualizar dados em uma forma tabular. Já a predição de desempenho dos estudantes, apresentou resultados satisfatórios, conseguindo em alguns cursos um *F1-score* acima de 0.85. No próximo capítulo, a análise continua por meio da visualização interativa de informações, onde, com a junção de técnicas visuais, aprendizado de máquina e interação com o usuário é possível extrair mais informações sobre os dados.

Capítulo 4

Visualização Exploratória de Dados Educacionais

Uma das formas de incluir o usuário em análises visuais é por meio da visualização interativa de dados, em que seus conhecimentos são fundamentais para a tomada de decisões [10]. Assim, para auxiliar a compreensão de dados educacionais, propõe-se a criação de uma ferramenta que auxilie o usuário nas tarefas de extração visual e interativa de dados multidimensionais. O restante deste capítulo é apresentado em quatro partes: A Seção 4.1 apresenta trabalhos semelhantes referentes a visualização exploratória; a Seção 4.2 descreve a metodologia utilizada para o desenvolvimento da ferramenta; a Seção 4.3 detalha o desenvolvimento de um protótipo juntamente com os resultados obtidos; por fim, as considerações finais sobre o capítulo são discutidas na Seção 4.4.

4.1 Revisão de Literatura

Existem vários trabalhos que englobam o tema de visualização exploratória de dados. Assim, para assistir o desenvolvimento deste projeto e descobrir exemplos de aplicações, nesta seção são apresentadas algumas pesquisas envolvendo técnicas de visualização interativa.

Naranjo et al. [73] desenvolve o *CloudTrail-Tracker*, um *Dashboard* integrado com a plataforma *Amazon Web Service* (AWS) para auxiliar o controle e o desempenho dos estudantes em disciplinas ministradas na nuvem por meio do AWS. O *Dashboard* tem como objetivos, mostrar informações sobre os recursos utilizados do AWS, detalhar informações específicas das atividades dos alunos e disponibilizar o progresso dos estudantes nas atividades. Todos estes, pertencentes a um intervalo de tempo definido pelo usuário. Além disso, o *CloudTrail-Tracker* separa os usuários entre três diferentes classes: professores, que podem verificar o desempenho de seus estudantes; alunos, que visualizam seu progresso nas atividades, seus planejamentos e quais atividades estão faltando; e administrador do sistema, onde são mostrados o consumo dos

recursos da plataforma em nuvem a cada 15 minutos. Além de permitir o acesso por meio de celulares, tablets e computadores, todas as informações são apresentadas por meio de gráficos interativos, onde o usuário pode filtrar o que ele deseja mostrar em sua tela durante o uso.

Para avaliar o funcionamento da ferramenta, foi verificada a satisfação dos estudantes que a utilizaram. A pesquisa envolvendo 64 alunos mostrou que 90% dos estudantes estão muito satisfeitos com a usabilidade e com as informações dispostas. Além disso, eles classificam a ferramenta como uma forma apropriada de integração com aulas ministradas na nuvem.

Fritze et al. [74] apresenta uma análise visual e interativa de dados curriculares dos estudantes de medicina da Alemanha. Fazendo parte do projeto intitulado *MERlin*, é proposta uma forma unificada de análise e medição de conhecimento por meio de uma plataforma online. Os dados dos currículos são classificados em quatro categorias que são usadas para gerar os gráficos interativos. Para obter as informações, o usuário filtra quais dados deseja acessar, podendo visualizar valores tanto de departamentos quanto de estudantes, e, a partir desta escolha, é gerado um gráfico de um tipo pré-selecionado que melhor representa a informação almejada. O projeto é considerado um sucesso, já sendo utilizado por 14 das 38 universidades de medicina da Alemanha, das quais seis já realizaram a total integração com o sistema.

Ltifi et al. [75] utiliza técnicas de mineração visuais dinâmicas para auxiliar na tomada de decisões médicas com respeito ao combate às infecções hospitalares nas unidades de tratamento intensivo. Neste estudo, é utilizado o conceito de dados temporais, que são definidos como dados que descrevem a evolução das características de um objeto ao decorrer de um intervalo de tempo. O método proposto investiga informações temporais interativamente, detecta relações importantes nos dados e identifica modelos relevantes a serem usados na tomada de decisão.

As técnicas de exploração visual são utilizadas em três camadas: manipulação de dados temporais, visualização temporal e gerenciamento de descoberta de conhecimento. A manipulação dos dados temporais se dá categorizando os dados gerais em relação ao seu *timestamp* em pares $< t, v >$ onde t denota a unidade de tempo (i.e. dia, hora ou minuto) e v denota a estrutura de dados representativa do modelo utilizado. A visualização temporal é feita transformando em uma adaptação de espaço-tempo como, por exemplo, uma linha do tempo contendo representações gráficas da estrutura de dados atrelada à cada valor temporal. O gerenciamento de descoberta de conhecimento, realizado pelo tomador de decisões, consiste na compreensão lógica do modelo visual e utilização do conhecimento proporcionado pela visualização temporal no processo de tomada de decisão. Mesmo em fases iniciais de desenvolvimento, o estudo mostra resultados satisfatórios, deixando testes experimentais a serem realizados em projetos futuros.

4.2 Metodologia Proposta

Para aumentar o conhecimento gerado a partir dos dados educacionais e aprofundar a análise de técnicas visuais com foco em necessidades específicas de informações, elaborou-se uma ferramenta de visualização exploratória interativa, na qual o usuário procura padrões e extrai informações sobre grupos específicos pertencentes ao dado com o auxílio de técnicas de aprendizado de máquina.

O funcionamento da ferramenta é descrito por meio da metodologia formada por cinco etapas sequenciais ilustrando o processo de extração da informação a partir de um conjunto de dados pronto para ser estudado, disposta na Figura 4.1. Sendo parte fundamental na visualização exploratória, o usuário está presente em todo o processo descrito da seguinte forma:

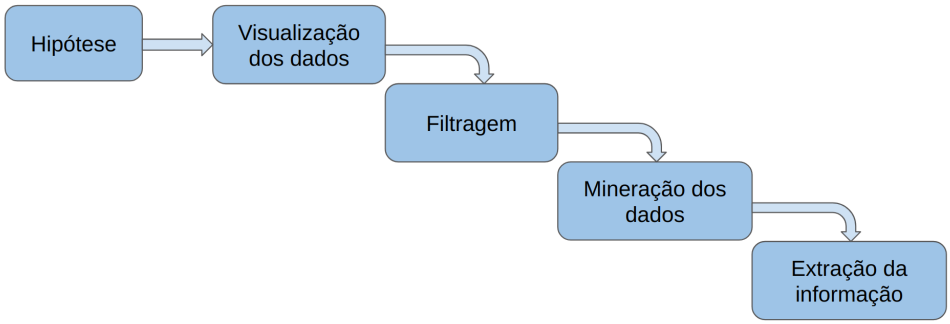


Figura 4.1: Metodologia proposta para a extração de conhecimento por mineração de dados.

1. **Hipótese**: Inicialmente, o usuário define a hipótese escolhendo o objetivo da visualização e quais a informação que deseja extrair dos dados disponíveis. Como, por exemplo, descobrir a média de faltas de determinado grupo de estudantes.

2. **Visualização dos dados**: Com a hipótese definida, é feita uma visualização aplicando-se alguma técnica baseada em redução de dimensionalidade, como o t-SNE ou o PCA, no conjunto de dados que o usuário tem interesse. O resultado dessa redução é mostrado em um *layout* bidimensional que possui uma ferramenta de seleção integrada.

3. **Filtragem**: Após a geração do gráfico, a filtragem é realizada de maneira interativa, onde o usuário seleciona os pontos de interesse no gráfico gerado na etapa anterior.

4. **Mineração dos dados**: O usuário ajusta interativamente a técnica de agrupamento e parâmetros associados à mesma, como por exemplo, o número de *clusters*. Em seguida, o agrupamento é aplicado ao subconjunto selecionado pelo próprio usuário na etapa de filtragem. Nesta etapa, são gerados um novo gráfico e uma tabela contendo os centróides dos *clusters*.

5. **Extração da informação**: Com o gráfico e os centróides obtidos na etapa de mineração de dados, o usuário consegue visualizar padrões no subconjunto selecionado e informações como, por exemplo, a média dos atributos pertencentes ao *cluster*.

4.3 Procedimentos e Resultados

Assim, desenvolveu-se um protótipo da ferramenta seguindo a metodologia, em que, para validar seu funcionamento e verificar suas utilidades, desenvolveu-se dois casos de usos utilizando, como dado de entrada, o subconjunto de alunos não cotistas do curso Ciência da Computação do CD3, descritos anteriormente na Seção 3.3.2.

Caso de Uso 1

Para o primeiro caso de uso, definiu-se a seguinte hipótese: " É possível estabelecer uma relação entre a forma de saída e as notas das matérias do primeiros semestre dos alunos de Ciência da Computação?".

Para responder a hipótese, a visualização t-SNE foi aplicada ao subconjunto selecionado, permitindo ao usuário visualizar bidimensionalmente todas as instâncias dos dados. A Figura 4.2 mostra o *layout* gerado a partir algoritmo t-SNE agrupando os alunos pelo *Estado de Graduação*.

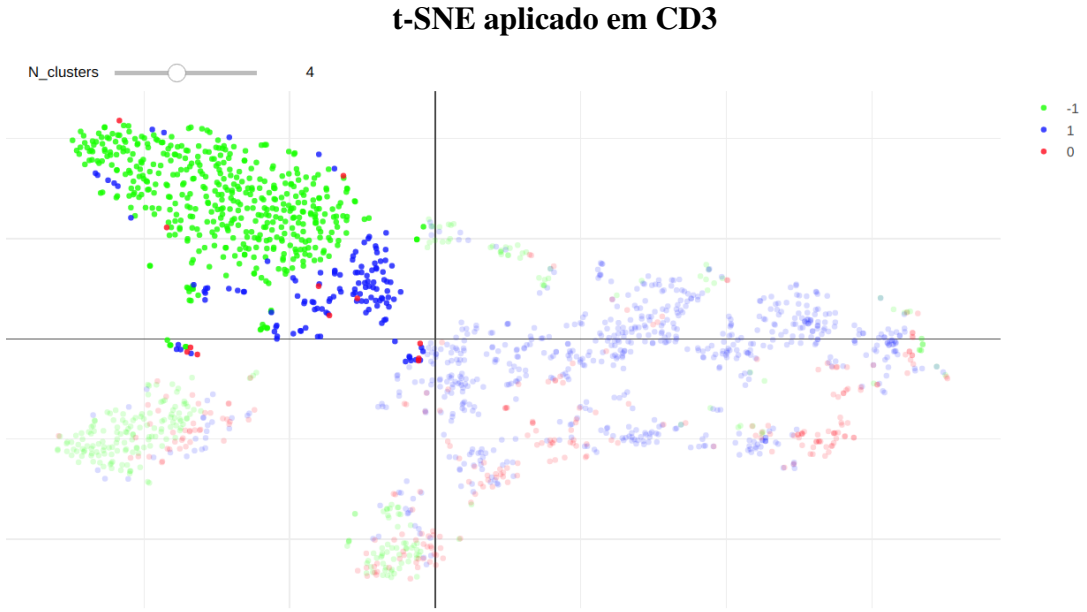


Figura 4.2: Gráfico do t-SNE com procedimento configurado para o evento de seleção de pontos em que os pontos destacados estão selecionados.

Identificada a parte do gráfico na qual se deseja extrair mais informações, o usuário escolhe o número de *clusters* que deseja criar e seleciona o subconjunto. As instâncias que pertencem ao subconjunto de interesse são então classificadas por meio do algoritmo de clusterização *K-means*, gerando assim o número de *clusters* que o usuário solicitou. A Figura 4.3 apresenta a imagem gerada após as etapas de filtragem e mineração de dados, ilustrando a disposição dos alunos em cada um dos quatro *clusters* que o usuário deseja criar.

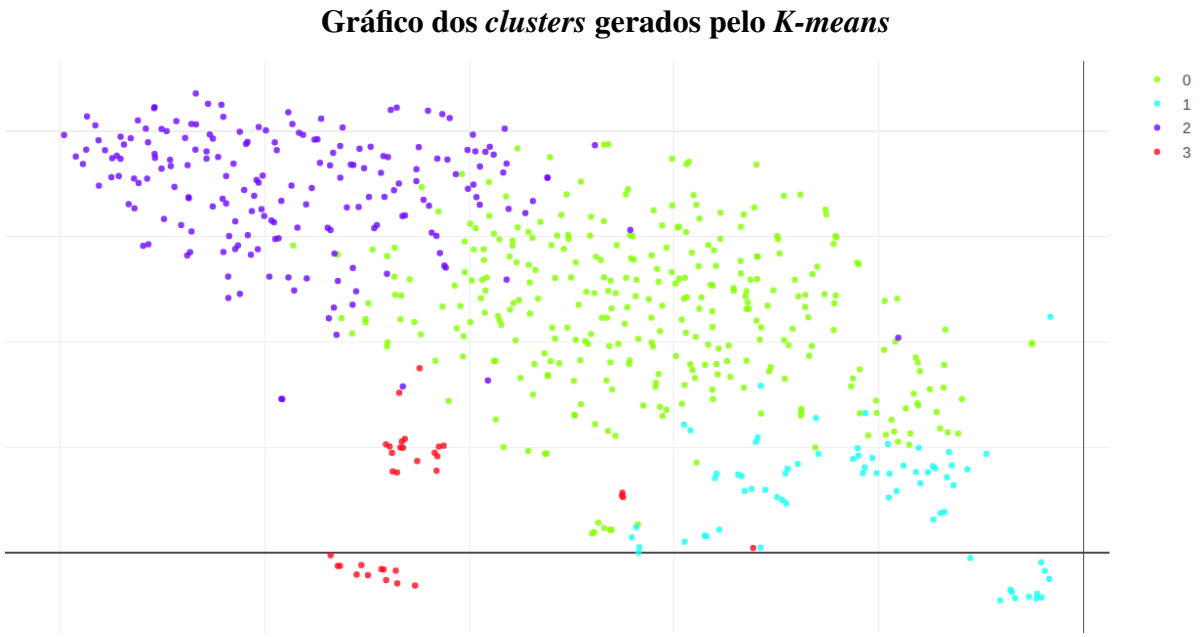


Figura 4.3: Gráfico dos *clusters* gerados no conjunto original a partir do subconjunto selecionado pelo usuário.

Além da geração do gráfico dos *clusters*, exibe-se também uma tabela contendo informações sobre o centróide de cada *cluster*, permitindo assim ao usuário verificar o valor médio dos atributos de cada grupo gerado. A Tabela 4.1 disponibiliza a nota média das seis disciplinas do primeiro semestre e o valor médio do atributo *Estado da graduação* dos alunos pertencentes a cada um dos quatro grupos gerados.

Estado da Graduação	116301	118001	118010	113034	140481	145971
-0.714733	3.530303	3.101880	3.302769	3.088296	3.821839	3.742424
0.857142	3.300000	2.922380	3.239285	2.895238	3.842857	3.500000
-0.884210	4.177192	3.876315	3.848245	3.994736	4.228947	4.107894
-0.181818	3.858585	3.272727	3.606060	2.787878	-0.868686	3.515151

Tabela 4.1: Centróides calculados dos *clusters* mostrados na Figura 4.3

A partir dos resultados apresentados na Tabela 4.1, o usuário pode extrair informações para responder a hipótese. No caso, nota-se que no conjunto de estudantes selecionados, a média das notas dos *clusters* com maior porcentagem de alunos formados (Estado da Graduação mais próximo a -1) tendem a ser maior comparado com os grupos com maior taxa de abandono (Estado da Graduação mais próximo a 1). Foralecendo assim a nossa hipótese de que os alunos com menor nota nas disciplinas do primeiro semestre tendem a abandonar o curso.

Caso de Uso 2

Para o segundo caso de uso, definiu-se a seguinte hipótese: "Os alunos que possuem menores notas nas matérias de Cálculo e Física possuem maior chance de abandonar o curso?".

Diferentemente do primeiro caso de uso, como técnica de redução de dimensionalidade foi aplicado o PCA, gerando um layout diferente da apresentada anteriormente na Figura 4.2. Assim, a Figura 4.4 representa um *layout* gerado a partir do algoritmo PCA agrupado pelo *Estado de Graduação* dos estudantes.

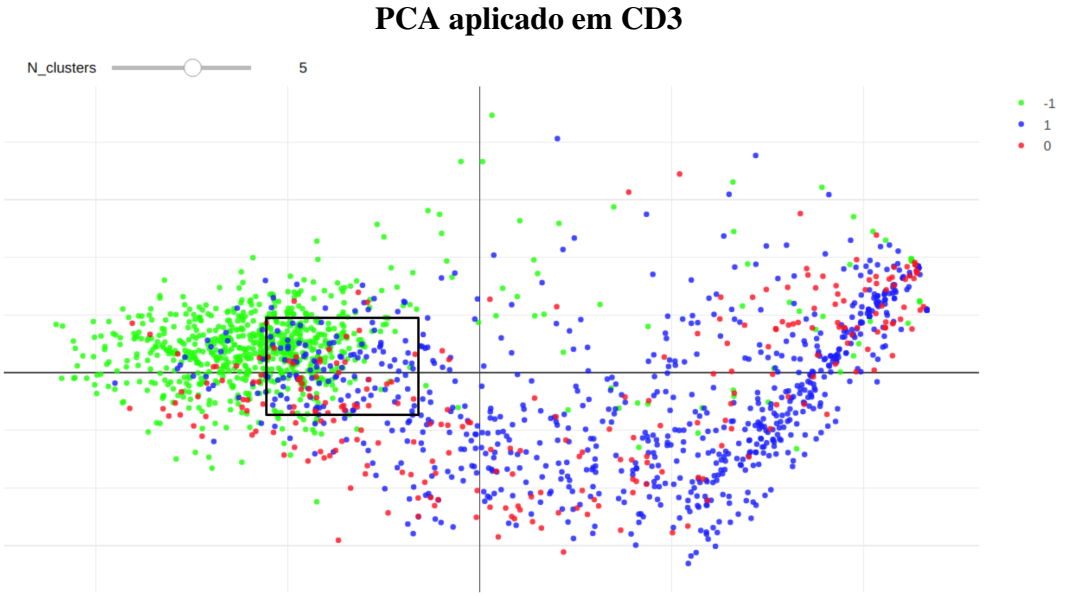


Figura 4.4: Gráfico do PCA com procedimento configurado para o evento de seleção de pontos.

A partir da visualização do gráfico, o usuário define a quantidade de *clusters* e seleciona as instâncias que deseja agrupar. Novamente, utilizou-se o algoritmo *K-means* no processo, desta vez gerando cinco grupos de alunos. A Figura 4.5 apresenta o gráfico gerado a partir do agrupamento separado nos cinco grupos almejados.

Gráfico dos *clusters* gerados pelo *K-means*

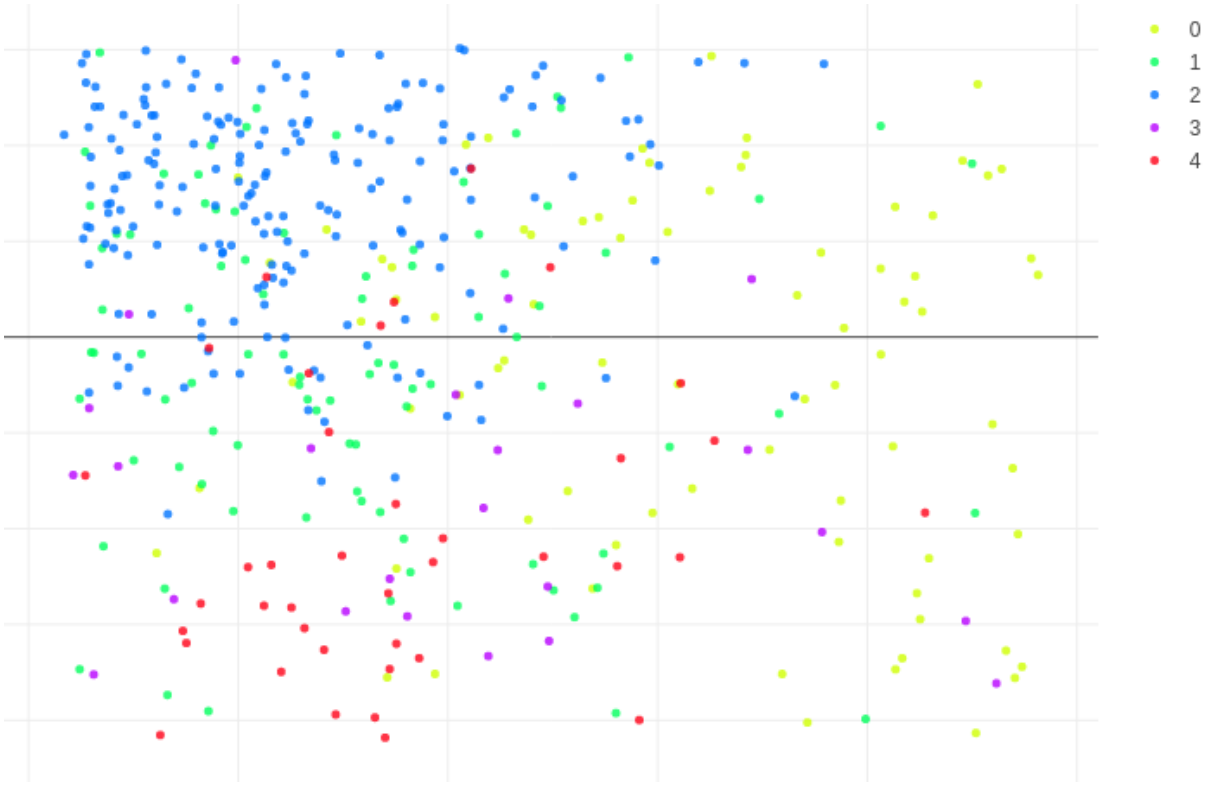


Figura 4.5: Gráfico dos *clusters* gerados no conjunto original a partir do subconjunto selecionado pelo usuário.

Assim como no caso anterior, juntamente com a visualização gráfica dos grupos, é gerada uma tabela contendo os centróides dos *clusters*. A Tabela 4.2 apresenta a nota média nas disciplinas de Cálculo e de Física juntamente com o *Estado de Graduação* dos cinco grupos.

Estado da Graduação	113034	113042	113051	118001	118028	118044
0.873417	2.905063	2.665400	0.916033	3.019831	2.821940	1.590717
-0.274725	2.677655	2.770146	2.218864	3.020146	2.948717	2.529304
-0.797927	3.091537	2.913471	2.250431	3.222797	3.089810	2.737046
-0.041666	2.965277	2.888888	2.315972	3.131944	2.9375	2.520833
-0.307692	2.782051	2.722222	2.632478	2.884615	2.935897	2.717948

Tabela 4.2: Centróides calculados dos *clusters* mostrados na Figura 4.5

A partir das informações extraídas no processo, nota-se que os grupos com maior proporção de alunos formados (Estado de Graduação mais próximo de -1) apresenta as melhores notas nas disciplinas. Porém, o grupo com maior taxa de abandono não possui as piores notas nas

disciplinas 113034 e 118001, contradizendo a hipótese de que os alunos que possuem menor nota em Cálculo e Física possuem maior chance de abandonar o curso.

4.4 Considerações Finais

Neste capítulo, apresentou-se uma metodologia de extração interativa de informações juntamente com a implementação de um protótipo desta metodologia, possibilitando assim, um maior controle ao usuário sobre as informações que se deseja estudar. O protótipo mostrou-se capaz de classificar os dados selecionados a partir de *clusters* e extrair deles as informações almejadas. A seguir, no Capítulo 5, serão apresentadas as conclusões finais sobre os três objetivos desta pesquisa juntamente com alguns trabalhos futuros sugeridos.

Capítulo 5

Conclusão

O desenvolvimento deste trabalho se decorreu a partir do estudo de dados educacionais de estudantes dos cursos de tecnologia da Universidade de Brasília. Destes, foram definidos três objetivos principais: realizar uma comparação visual entre os alunos cotistas e não cotistas; criar um modelo de predição para identificar se os alunos abandonarão o curso; e desenvolver um método de extração de informação de forma interativa.

Comparando os estudantes, percebeu-se que os alunos não cotistas apresentam um desempenho melhor nas matérias iniciais e possuem uma menor porcentagem de evasão nos cursos. Adicionalmente, notou-se que em ambos os grupos existe uma maior taxa de reprovação nos primeiros semestres. Também verificou-se a visualização dos dados ao aplicar técnicas de redução de dimensionalidade, da qual, com a aplicação do PCA e do t-SNE, foi possível identificar uma tendência de separação entre os alunos que formam ou abandonam seus respectivos cursos.

Para realizar a classificação dos alunos, com o intuito de descobrir qual algoritmo teria o melhor rendimento, utilizou-se o *K-NN* e o *Gradient Boosting*. Como entrada para o modelo, foram escolhidas as notas médias do aluno nas matérias referentes aos primeiros três semestres dos cursos. O *Gradient Boosting* demonstrou mais eficiência, atingindo, na métrica *F1-score*, 0.87 e 0.94 para os alunos de Ciência da computação e Engenharia da computação, respectivamente. Mostrando que as notas nas matérias do começo do curso exercem forte influência no abandono dos alunos.

No processo de visualização interativa, desenvolveu-se um protótipo capaz de proporcionar ao usuário maior controle sobre as informações que se deseja extrair. De forma interativa, o usuário pode selecionar o grupo de alunos representados em um gráfico e agrupá-los em *clusters*. A partir destes grupos criados, é possível encontrar padrões entre os estudantes e verificar dados como a média de notas das matérias ou a porcentagem de alunos que formaram ou abandonaram o curso.

Espera-se que as informações extraídas em conjunto com a protótipo desenvolvido neste projeto possa ajudar os professores e o departamento gerando uma maior compreensão da situ-

ação dos estudantes, auxiliando a tomada de decisões em como ministrar disciplinas, assistir os alunos e melhorar a qualidade do ensino do departamento. Utilizando, por exemplo, o modelo de predição para encontrar os alunos com maior chance de abandonar o curso e realizar um acompanhamento em conjunto com a coordenação do departamento ao qual o aluno pertence.

5.1 Trabalhos Futuros

Mesmo obtendo resultados satisfatórios no estudos, percebeu-se que ainda existem mais alunos cotistas cursando os cursos do que formados. Assim, sugere-se que passados alguns anos, repita-se a comparação entre os alunos cotistas e não cotistas para conseguir dados mais precisos e poder avaliar, a partir dos informações obtidas neste estudo, a evolução dos estudantes.

Quanto a predição de abandono, foram utilizadas somente duas técnicas de aprendizagem de máquina, assim, propõe-se a implementação de outras, como por exemplo redes neurais. Além disso, sugere-se somente a utilização de dados produzidos a partir dos primeiros dois semestres do curso para prever a evasão com mais antecedência proporcionando mais tempo de ação para os professores e ao departamento.

Sugere-se também, para o desenvolvimento da visualização interativa, a utilização de outros algoritmos de agrupamento e de redução de dimensionalidade deixando ao usuário a escolha da melhor técnica a ser utilizada. além disso, a integração de um modelo de predição com a ferramenta pode gerar informações úteis mostrando por exemplo a probabilidade de se formar de cada estudante do grupo selecionado.

Referências

[1] Baker, Ryan S: *Educational data mining: An advance for intelligent systems in education*. IEEE Intelligent systems, 29(3):78–82, 2014. 1

[2] Romero, Cristóbal e Sebastián Ventura: *Educational data mining: a review of the state of the art*. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6):601–618, 2010. 1, 2

[3] Al-Radaideh, Qasem A, Emad M Al-Shawakfa e Mustafa I Al-Najjar: *Mining student data using decision trees*. Em *International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan*, 2006. 1

[4] Education Statistics, Washington's Department of Education: National Center for: *Forum guide to education data privacy*. 2016. 1

[5] Education Reform, The Glossary of: *Student-level data*. `https://www.edglossary.org/student-level-data/`, acesso em 2019-05-15. 1

[6] Santos Baggi, Cristiane Aparecida dos e Doraci Alves Lopes: *Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica*. Avaliação: Revista da Avaliação da Educação Superior, 16(2), 2011. 2

[7] David, Lamartine Moreira Lima e Carlos Dias Chaym: *Evasão universitária: Um modelo para diagnóstico e gerenciamento de instituições de ensino superior*. Revista de Administração IMED, 9(1):167–186, 2019. 2

[8] Paulo Lima Junior, Nilce Santos de Melo e Mauro Rabelo: *Estatísticas de trajetória dos alunos de graduação – instituto de ciências exatas(ie)*. 2016. 2

[9] Dewan, Pauline: *Words versus pictures: leveraging the research on visual communication*. Partnership: the Canadian Journal of Library and Information Practice and Research, 10(1), 2015. 2

[10] Keim, Daniel A: *Information visualization and visual data mining*. IEEE transactions on Visualization and Computer Graphics, 8(1):1–8, 2002. 2, 8, 40

[11] Beck, Joseph, Ryan Baker, Albert T. Corbett, Judy Kay, Diane Litman, Antonija Mitrovic e Steve Ritter: *Workshop on analyzing student-tutor interaction logs to improve educational outcomes*. página 909, janeiro 2004. 2

[12] Bresfelean, Vasile Paul, Mihaela Bresfelean, Nicolae Ghisoiu e Calin Adrian Comes: *Determining students' academic failure profile founded on data mining methods*. Em *ITI 2008-30th International Conference on Information Technology Interfaces*, páginas 317–322. IEEE, 2008. 2

[13] Dekker, Gerben W, Mykola Pechenizkiy e Jan M Vleeshouwers: *Predicting students drop out: A case study.* International Working Group on Educational Data Mining, 2009. 2

[14] Delavari, Naeimeh, Somnuk Phon-Amnuaisuk e Mohammad Reza Beikzadeh: *Data mining application in higher learning institutions*. 2008. 2

[15] Avouris, Nikolaos, Vassilis Komis, Georgios Fiotakis, Meletis Margaritis e Eleni Voyiatzaki: *Logging of fingertip actions is not enough for analysis of learning activities*. Em *12th International Conference on Artificial Intelligence in Education, AIED 05 Workshop 1: Usage analysis in learning systems*, páginas 1–8, 2005. 2

[16] Chan, Chien Chung: *A framework for assessing usage of web-based e-learning systems*. Em *Second International Conference on Innovative Computing, Informatio and Control (ICICIC 2007)*, páginas 147–147. IEEE, 2007. 2

[17] Chen, Gwo Dong, Chen Chung Liu, Kuo Liang Ou e Baw Jhiune Liu: *Discovering decision knowledge from web log portfolio for managing classroom processes by applying decision tree and data cube technology*. Journal of Educational Computing Research, 23(3):305–332, 2000. 2

[18] Chen, Chih Ming, Ling Jiun Duh e Chao Yu Liu: *A personalized courseware recommendation system based on fuzzy item response theory*. Em *IEEE International Conference on e-Technology, e-Commerce and e-Service, 2004. EEE'04. 2004*, páginas 305–308. IEEE, 2004. 2

[19] Baruque, Cássia Blondet, Marília A Amaral, Alexandre Barcellos, João Carlos da Silva Freitas e Carlos Juliano Longo: *Analysing users' access logs in moodle to improve e learning*. Em *Proceedings of the 2007 Euro American conference on Telematics and information systems*, página 72. ACM, 2007. 2

[20] Chanchary, Farah Habib, Indrani Haque e Md Saifuddin Khalid: *Web usage mining to evaluate the transfer of learning in a web-based learning environment*. Em *First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008)*, páginas 249–253. IEEE, 2008. 2

[21] Chang, Yi Chun, Wen Yan Kao, Chih Ping Chu e Chiung Hui Chiu: *A learning style classification mechanism for e-learning*. Computers & Education, 53(2):273–285, 2009. 2

[22] Fisher, Ronald A: *The use of multiple measurements in taxonomic problems*. Annals of eugenics, 7(2):179–188, 1936. 4

[23] De Oliveira, MC Ferreira e Haim Levkowitz: *From visual data exploration to visual data mining: a survey*. IEEE Transactions on Visualization and Computer Graphics, 9(3):378–394, 2003. 5

[24] Statistics, PennState Department of: *Measures of similarity and dissimilarity*. https://newonlinecourses.science.psu.edu/stat508/lesson/1b/1b.2/1b.2.1, acesso em 2019-05-15. 5, 6

[25] Muniz, Sérgio Ricardo: *Introdução à análise estatística de medidas*. https://edisciplinas.usp.br/pluginfile.php/4394902/mod_resource/content/0/plc0016_14.pdf, acesso em 2019-05-15. 6

[26] Rice, John A: *Mathematical statistics and data analysis*. Cengage Learning, 2006. 7

[27] Lee Rodgers, Joseph e W Alan Nicewander: *Thirteen ways to look at the correlation coefficient*. The American Statistician, 42(1):59–66, 1988. 7

[28] Gradoni, Gabriele, Valter Mariani Primiani e Franco Moglie: *Reverberation chamber as a multivariate process: Fdtd evaluation of correlation matrix and independent positions*. Progress In Electromagnetics Research, 133:217–234, 2013. 7

[29] Khan, Muzammil e Sarwar Shah Khan: *Data and information visualization methods, and interactive mechanisms: A survey*. International Journal of Computer Applications, 34(1):1–14, 2011. 8, 9, 10

[30] Wang, Lidong, Guanghui Wang e Cheryl Ann Alexander: *Big data and visualization: methods, challenges and technology progress*. Digital Technologies, 1(1):33–38, 2015. 8

[31] Zhao, Shilin, Yan Guo, Quanhu Sheng e Yu Shyr: *Advanced heat map and clustering analysis using heatmap3*. BioMed research international, 2014, 2014. 11

[32] Pleil, Joachim D, Matthew A Stiegel, Michael C Madden e Jon R Sobus: *Heat map visualization of complex environmental and biomarker measurements*. Chemosphere, 84(5):716–723, 2011. 11

[33] Chen, Jim X e Shuangbao Wang: *Data visualization: parallel coordinates and dimension reduction*. Computing in Science & Engineering, 3(5):110–113, 2001. 12

[34] Bellman, Richard E: *Adaptive control processes: a guided tour*, volume 2045. Princeton university press, 2015. 13

[35] Paulovich, Fernando Vieira: *Mapeamento de dados multi-dimensionais-integrando mineração e visualização*. Tese de Doutoramento, Universidade de São Paulo, 2008. 14

[36] Van Der Maaten, Laurens, Eric Postma e Jaap Van den Herik: *Dimensionality reduction: a comparative*. J Mach Learn Res, 10(66-71):13, 2009. 14

[37] Pezzotti, Nicola: *Dimensionality-reduction algorithms for progressive visual analytics*. 2019. 14

[38] Jolliffe, Ian: *Principal component analysis*. Springer, 2011. 14

[39] Van Der Maaten, Laurens: *Accelerating t-sne using tree-based algorithms*. The Journal of Machine Learning Research, 15(1):3221–3245, 2014. 15, 16

[40] Raykar, Vikas Chandrakant e Ramani Duraiswami: *Fast optimal bandwidth selection for kernel density estimation*. Em *Proceedings of the 2006 SIAM International Conference on Data Mining*, páginas 524–528. SIAM, 2006. 16

[41] Broda, Simon A, Jochen Krause e Marc S Paolella: *Approximating expected shortfall for heavy-tailed distributions*. Econometrics and statistics, 8:184–203, 2018. 16

[42] Bernardi, Mauro, Valeria Bignozzi e Lea Petrella: *On the lp-quantiles for the student t distribution*. Statistics & Probability Letters, 128:77–83, 2017. 16

[43] Agahi, Hamzeh: *A modified kullback–leibler divergence for non-additive measures based on choquet integral*. Fuzzy Sets and Systems, 2019. 16

[44] Maaten, Laurens van der e Geoffrey Hinton: *Visualizing data using t-sne*. Journal of machine learning research, 9(Nov):2579–2605, 2008. 16

[45] Russell, Stuart J e Peter Norvig: *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016. 18

[46] Tseng, Kuo Kun, Regina Fang Ying Lin, Hongfu Zhou, Kevin Jati Kurniajaya e Qianyu Li: *Price prediction of e-commerce products through internet sentiment analysis*. Electronic Commerce Research, 18(1):65–88, 2018. 18

[47] Komura, Daisuke e Shumpei Ishikawa: *Machine learning methods for histopathological image analysis*. Computational and structural biotechnology journal, 16:34–42, 2018. 18

[48] Stilgoe, Jack: *Machine learning, social learning and the governance of self-driving cars*. Social studies of science, 48(1):25–56, 2018. 18

[49] Han, Jae Hyun, Kang Min Bae, Seong Kwang Hong, Hyunsin Park, Jun Hyuk Kwak, Hee Seung Wang, Daniel Juhyung Joe, Jung Hwan Park, Young Hoon Jung, Shin Hur *et al.*: *Machine learning-based self-powered acoustic sensor for speaker recognition*. Nano energy, 53:658–665, 2018. 18

[50] Islam, Md Rafiqul, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M Kamal, Hua Wang e Anwaar Ulhaq: *Depression detection from social network data using machine learning techniques*. Health information science and systems, 6(1):8, 2018. 18

[51] Berland, Matthew, Ryan S Baker e Paulo Blikstein: *Educational data mining and learning analytics: Applications to constructionist research*. Technology, Knowledge and Learning, 19(1-2):205–220, 2014. 18

[52] Jain, Anil K, Richard C Dubes *et al.*: *Algorithms for clustering data*, volume 6. Prentice hall Englewood Cliffs, 1988. 18

[53] Selim, Shokri Z e Mohamed A Ismail: *K-means-type algorithms: A generalized convergence theorem and characterization of local optimality*. IEEE Transactions on pattern analysis and machine intelligence, (1):81–87, 1984. 18, 19

[54] Bobrowski, Leon e James C Bezdek: *C-means clustering with the l/sub l/and l/sub infinity/norms*. IEEE Transactions on Systems, Man, and Cybernetics, 21(3):545–554, 1991. 18

[55] Zhou, Pei Yuan e Keith CC Chan: *A model-based multivariate time series clustering algorithm*. Em *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, páginas 805–817. Springer, 2014. 19

[56] Huang, Zhexue: *Extensions to the k-means algorithm for clustering large data sets with categorical values*. Data mining and knowledge discovery, 2(3):283–304, 1998. 19

[57] Cover, Thomas M, Peter E Hart *et al.*: *Nearest neighbor pattern classification*. IEEE transactions on information theory, 13(1):21–27, 1967. 19

[58] Friedman, Jerome H: *Stochastic gradient boosting*. Computational statistics & data analysis, 38(4):367–378, 2002. 20, 21

[59] Friedman, Jerome H: *Greedy function approximation: a gradient boosting machine*. Annals of statistics, páginas 1189–1232, 2001. 20

[60] Bonfim, Edmar Ferreira Souto Mourão: *Avaliação do rendimento e evasão de alunos cotistas e não cotistas da universidade de brasília*. 2014. 22, 24

[61] Saa, Amjad Abu *et al.*: *Educational data mining & students' performance prediction*. International Journal of Advanced Computer Science and Applications, 7(5):212–220, 2016. 22

[62] Fernandes, Eduardo, Maristela Holanda, Marcio Victorino, Vinicius Borges, Rommel Carvalho e Gustavo Van Erven: *Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil*. Journal of Business Research, 94:335–343, 2019. 23, 38

[63] Costa, Fellipe, Anderson Rufino dos Santos Silva, Daniel Miranda de Brito e Thaís Gaudêncio do Rêgo: *Predição de sucesso de estudantes cotistas utilizando algoritmos de classificação*. Em *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 26, página 997, 2015. 23

[64] Dário, Amalia Borges *et al.*: *Avaliação do desempenho acadêmico e da evasão entre discentes cotistas e não cotistas*. 2017. 23

[65] Luiza A. Hansen, Lucas M. Chagas: *Análise visual dos dados educacionais voltada para o estudo de gênero nos cursos de computação da universidade de brasília*. 2018. 37

[66] Papernot, Nicolas e Patrick McDaniel: *Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning*. arXiv preprint arXiv:1803.04765, 2018. 38

[67] Wu, Xueyan, Jiquan Yang e Shuihua Wang: *Tea category identification based on optimal wavelet entropy and weighted k-nearest neighbors algorithm*. Multimedia Tools and Applications, 77(3):3745–3759, 2018. 38

[68] Müller, Philipp, Katri Salminen, Ville Nieminen, Anton Kontunen, Markus Karjalainen, Poika Isokoski, Jussi Rantala, Mariaana Savia, Jari Väliaho, Pasi Kallio *et al.*: *Scent classification by k nearest neighbors using ion-mobility spectrometry measurements*. Expert Systems with Applications, 115:593–606, 2019. 38

[69] Chen, Xing, Li Huang, Di Xie e Qi Zhao: *Egbmmda: extreme gradient boosting machine for mirna-disease association prediction*. Cell death & disease, 9(1):3, 2018. 38

[70] Rao, Haidi, Xianzhang Shi, Ahoussou Kouassi Rodrigue, Juanjuan Feng, Yingchun Xia, Mohamed Elhoseny, Xiaohui Yuan e Lichuan Gu: *Feature selection based on artificial bee colony and gradient boosting decision tree*. Applied Soft Computing, 74:634–642, 2019. 38

[71] Ling, Hao, Chunxiang Qian, Wence Kang, Chengyao Liang e Huaicheng Chen: *Combination of support vector machine and k-fold cross validation to predict compressive strength of concrete in marine environment*. Construction and Building Materials, 206:355–363, 2019. 38

[72] Fujino, Akinori, Hideki Isozaki e Jun Suzuki: *Multi-label text categorization with model combination based on f1-score maximization*. Em *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008. 38

[73] Naranjo, Diana M, José R Prieto, Germán Moltó e Amanda Calatrava: *A visual dashboard to track learning analytics for educational cloud computing*. Sensors, 19(13):2952, 2019. 40

[74] Fritze, Olaf, Maria Lammerding-Koeppel, Martin Boeker, Elisabeth Narciss, Annette Wosnik, Stephan Zipfel e Jan Griewatz: *Boosting competence-orientation in undergraduate medical education–a web-based tool linking curricular mapping and visual analytics*. Medical teacher, páginas 1–11, 2018. 41

[75] Ltifi, Hela, Emna Benmohamed, Christophe Kolski e Mounir Ben Ayed: *Enhanced visual data mining process for dynamic decision-making*. Knowledge-Based Systems, 112:166–181, 2016. 41

Anexo I

Tabela do Código da Disciplina com o Respectivo Nome das Matérias Citadas

Código	**Nome da Disciplina**
116301	Computação Básica
113042	Cálculo 2
118001	Física 1
113093	Introdução a Álgebra Linear
167959	Fundamentos de Redes
168769	Mecânica 1
117366	Lógica Computacional 1
113476	Algoritmos e Programação de Computadores
110141	Tópicos Especiais em Programação
118010	Física Experimental 1
113034	Cálculo 1
140481	Leitura e Produção de Texto
145971	Inglês Instrumental 1
113051	Cálculo 3
118028	Física 2
118044	Física 3

Tabela I.1: Tabela do código da disciplina com o respectivo nome das matérias citadas.